The Genetics of Cognitively Healthy Centenarians

Tesi, Niccolò

2021

The Genetics of Cognitively Healthy Centenarians

Niccolò Tesi

The Genetics of Cognitively Healthy Centenarians

Niccolo' Tesi

100-plus Study ~ Alzheimer Center Amsterdam ~ Amsterdam ~ The Netherlands

github.com/TesiNicco

The work described in this thesis was founded by Stichting Alzheimer Nederland and Stichting VUmc fonds.
ISBN: 978-94-6423-450-3
Layout: Niccolò Tesi

Cover: *Flipping over the DNA*
Cover design: Arifhusni ~ *fiverr*

Final release, September 2021

VRIJE UNIVERSITEIT

The Genetics of Cognitively Healthy Centenarians

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor
aan de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geneeskunde
op dinsdag 28 september 2021 om 11.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Niccolo' Tesi

geboren te Firenze, Italië

promotoren:	prof. dr. W.M. van der Flier prof. dr. ir. M.J.T. Reinders
copromotor:	dr. H. Holstege
promotiecommissie:	prof. dr. D. Posthuma prof. dr. K. Sleegers prof. dr J. Hardy prof. dr. M. Verhage dr. A. Ruiz Laza

Contents

1. General Introduction

1.1 Cognitively healthy centenarians

One important accomplishment of humankind is the extension of the average life expectancy. Worldwide, this phenomenon has shown a remarkable linear increase over the last two centuries,[1] and by 2050 there will be 3.2 million of centenarians in the world.[2] However, a consequence of an aged population is the increased prevalence of age-related diseases.[3] Therefore, an increasing fraction of individuals will spend part of their old age in disability or dependence on others. A major contributor to poor health at old age is cognitive decline and dementia, of which Alzheimer's disease (AD) is the most common type. [4, 5] However, dementia is not an inevitable consequence of aging: in fact, a small proportion of the population (<0.1%) reaches at least 100 years of age while maintaining a high level of cognitive and physical functions, so-called cognitively healthy centenarians.[6, 7] This raises questions as to what extent these centenarians have exceptional features that protect or delay the onset of dementia and other age-related diseases, and to what extent genetic factors are involved. To find an answer to these questions, the 100-plus Study was initiated: a prospective cohort study that aims at unraveling the environmental and genetic factors that are associated with becoming a cognitively healthy centenarian.[6]

1.2 Epidemiology and genetics of Alzheimer's disease

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by the loss of cognitive functions, ultimately leading to loss of independence, and death.[5, 8] In the aged Western populations it is currently one of the most prevalent diseases and poses a huge burden on patients, their families, and society.[5, 4] Currently, there is no effective treatment to prevent or to slow AD progression.[5, 8] The prevalence of AD increases exponentially with age: while the disease is rare before the age of 65 years (early-onset AD, EOAD), the more common form of the disease, late onset AD (LOAD, age at onset >65 years), reaches ~40% per year at 100 years of age.[9] Next to aging, genetic factors play an important role: in fact, twin studies indicated that the heritability of the common form of AD ranges between 60-80%.[10] The strongest genetic risk factor for AD is *APOE* genotype, which was identified in the early 1990s through linkage studies, and in the Caucasian population determines up to 30% of the genetic risk of AD.[11, 12, 13] The *APOE* genotype for each individual is determined by the combination of two out of three alleles ($\epsilon 2$, $\epsilon 3$, and $\epsilon 4$), that make up the six possible genotypes ($\epsilon 2/\epsilon 2$, $\epsilon 2/\epsilon 3$, $\epsilon 2/\epsilon 4$, $\epsilon 3/\epsilon 3$, $\epsilon 3/\epsilon 4$ and $\epsilon 4/\epsilon 4$). The

ϵ3 allele is the most frequent in the population (77%), followed by the ϵ4 allele (15%) and lastly the ϵ2 allele (8%). In terms of AD risk, the ϵ4 allele increases AD susceptibility, while the ϵ3 is neutral and the ϵ2 allele has a protective effect against AD. Compared to carrying no ϵ4 allele, carrying one copy of the ϵ4 allele (*i.e.* ϵ3/ϵ4 heterozygous genotype) increases AD-risk by approximately 3-5 fold, while AD-risk is increased to 15-30 fold in individuals who carry two copies of the ϵ4 allele (*i.e.* ϵ4/ϵ4 homozygous genotype).[11, 12, 13, 14] With the development of genotyping arrays in the early 2000s, genome-wide association studies (GWAS) became possible, leading to the identification of additional genetic risk variants for LOAD.[15] Typically, in GWAS the frequency of genetic variants is compared between a group of individuals that manifest the phenotype of interest (cases) and a group of individuals in which the phenotype of interest is absent (controls).[15] Unlike linkage and family studies, typically hypothesis-driven, GWAS do not require any prior knowledge and thus have the potential to reveal new genetic discriminants of a given phenotype.[16] Additionally, the continuous development of reference panels comprising tens of thousands of individuals and next-generation imputation strategies have drastically improved the number of genetic variants that can be analyzed, whilst simultaneously reducing genotyping costs.[17, 18] Successive waves of GWAS of AD with an increasing number of individuals have been performed, and the number of variants identified to influence the genetic risk to develop AD has steeply increased.[19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29] Today, about 70 genetic variants in addition to the *APOE* variants have been associated with a slight modification of the risk of AD from GWAS. [30] Besides being a powerful instrument to delineate the genetic landscape of AD, GWAS studies have been pivotal to understand the molecular events that are associated with AD pathology (Figure 1.1). The amyloid hypothesis, which was proposed during the first years of AD research, is the commonly adopted theory to explain AD development. According to this theory, the imbalance between β-amyloid production and clearance (*i.e.* β-amyloid metabolism) is at the basis of the molecular cascade that lead to neuronal loss and ultimately to cognitive decline (Figure 1.1).[31] However, drugs targeting β-amyloid have as of yet been insufficiently effective, as the decrease in the amount of β-amyloid in the brain did not reduce the rate of disease progression and cognitive decline. This has led, together with a better understanding of the genetic factors that are associated with AD, to an evolution of the traditional β-amyloid theory to encompass more complex disease aspects.[32] For example, part of the current view of the etiology of AD is that the dysregulation of the endo-

lysosomal trafficking system and the immune response is a major causal pathway, and that AD is not just a consequence of β-amyloid metabolism (Figure 1.1).[33, 34] However, the extent to which different pathways associated with AD overlap and contribute to the total risk to develop the disease is mostly unknown.

1.3 Genetic factors underlying AD and longevity

Given the high prevalence of AD at old ages, and the importance of genetic factors for AD, cognitively healthy centenarians are exceptional individuals to study. Theoretically, it would be expected that genetic variants that are associated with an increased risk to develop AD should have a negative effect on longevity, as AD is associated with increased mortality. Therefore, the frequency of these variants in extremely old and healthy individuals is expected to be lower than in the general population. In fact, the largest genetic risk factor for AD, *APOE*, is also the largest genetic factor known to influence human longevity.[35] Surprisingly, only the *APOE* genotype is associated with both the genetics of AD and longevity, with the $\epsilon 4$ allele that increases AD risk and is associated with reduced longevity, and the $\epsilon 2$ allele which decreases AD risk and promote longer lifespan.[36] Such a small overlap between the genetics of AD and longevity may be attributable to several reasons: first, across different populations and cohorts the genetics of AD may be more homogeneous than the genetics of longevity; second, the case-control approach typically applied in studies investigating the genetics of AD cannot be robustly applied in studies investigating the genetics of longevity. For example, in studies of longevity it is unclear which individuals should be considered as cases and/or as controls while for AD affected individuals may be compared to (age- and population-matched) non-affected individuals; third, given that longevity is the result of resisting or delaying all deadly diseases, the effect on longevity of genetic variants associated with one specific disease may be relatively small, such that a large number of individuals of extreme age need to be compared to identify associated genetic variants, and that is not always feasible.[16] Despite great interest in understanding the relationship between cognitive decline due to AD and extreme longevity with retained cognitive health, the extent to which cognitively healthy super-agers are genetically protected against AD is largely unknown, and therefore object of our investigations.

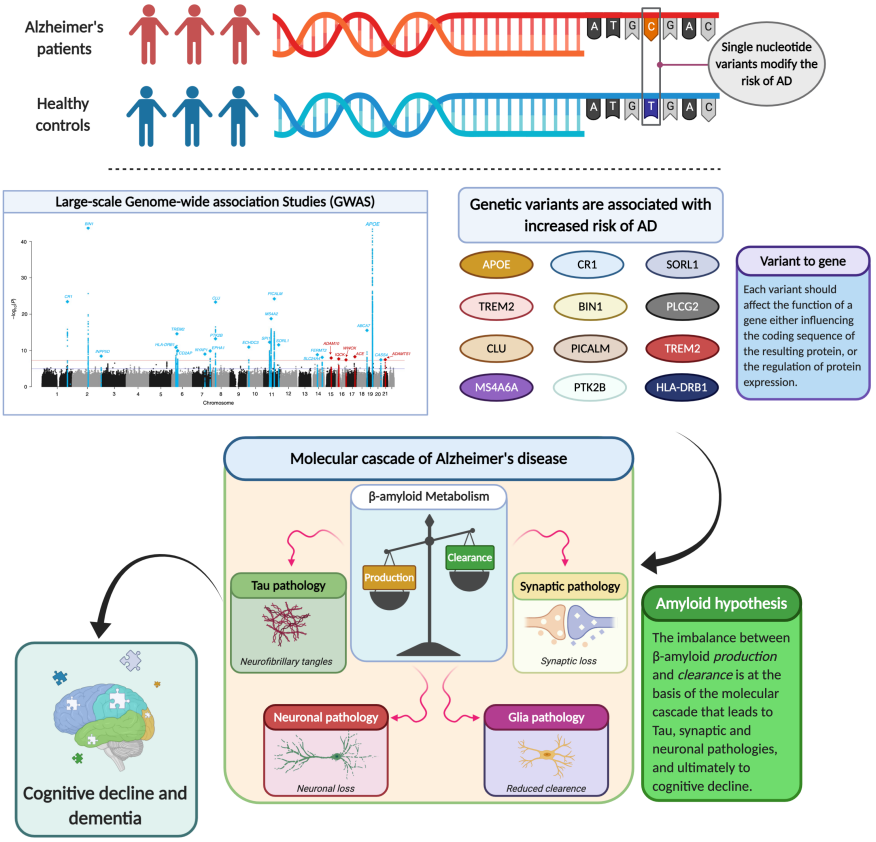


Figure 1.1: **Genetics underlying Alzheimer's disease.** The genetics factors at the basis of Alzheimer's disease are typically studied through Genome-Wide Association Studies (GWAS), where the frequency of genetic variants across the genome is compared between a sample of individuals with AD and a sample of individuals without AD. Genetic variants with a significant difference of frequency between AD patients and controls modify the risk to develop AD. Each genetic variant is thought to act at the gene level by regulating its expression or regulation. These genes, in turnm play a role in the molecular events that lead to AD. The commonly accepted theory explaining AD is based on the centrality of the β-amyloid metabolism. Specifically, the imbalance between β-amyloid production and clearence is thought to initiate the molecular cascade that eventually lead to cognitive decline and dementia.

1.4 Genetics of human longevity

Human longevity is one of the most complex phenotypes to study and it is influenced by environmental and genetic variables (Figure 1.2).[3] Long-lived individuals tend to cluster in families, which suggests that genetic factors play a role in determining extreme human longevity. In fact, although the heritability of lifespan up to ~70 years of age ranges only 10-25%, the heritability of becoming a centenarian raises up to ~50%.[37, 38, 39] This means that to reach higher ages we become increasingly dependent on the favorable genetic elements of our genomes. Ever since the GWAS-era started, many GWAS of longevity have been performed, trying to unravel the genetic architecture of extreme human aging and the relationship with age-related diseases.[40, 41, 42, 43, 44, 45, 46] As a result, a constellation of genetic variants has been associated with extended lifespan in independent studies. However, apart from *APOE* and few other candidates (*CDKN2B, ABO*), the replication of these genomic regions in independent studies has been challenging, in part due to heterogeneity in the study designs, methodologies, and populations. While not fully replicated, the genetic variants discovered thus far were known to associate with age-related diseases, including cardiovascular diseases and cancer, and with immunological and metabolic signatures, which are known hallmarks of aging (Figure 1.2).[42, 43, 46] Altogether, this suggests that an extended human lifespan is associated with a lower genetic risk of age-related diseases.[46][47][48] Given the uncertainties that are associated with the genetic factors associated with extreme longevity, it is of interest to determine the extent to which cognitively healthy agers are genetically predisposed to live longer, and whether this information can be used to predict overall survival.

1.5 GWAS limitations and alternative approaches

Despite the robustness of the GWAS approach to study complex polygenic traits such as longevity and AD, there are some limitations to this strategy. First, very large sample sizes are necessary to achieve sufficient statistical power: the power to detect significant associations is indeed a function of the sample size, the variant effect-size (how much the variant-frequency is different between cases and controls), and the significance threshold used (the evidence-level that there is a true difference between the variant-frequency in cases and controls).[16] Given that the effect-size of variants affecting complex traits are mostly small, and that the burden of multiple test correction is massive due to the high number of variants that are tested, it

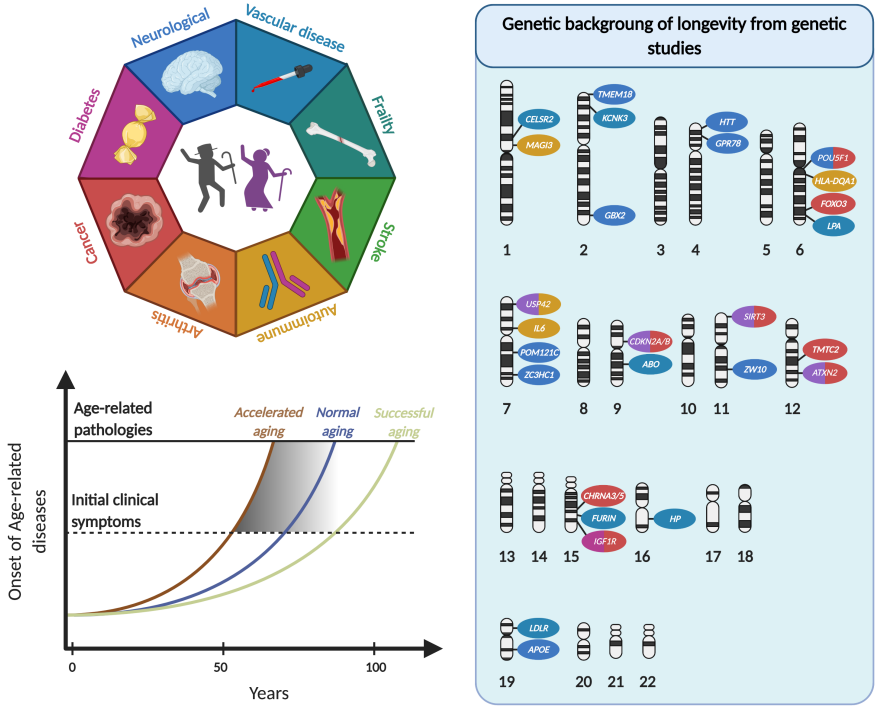


Figure 1.2: **Genetics underlying human longevity.** The genetic factors that influence human longevity are thought to resemble those influencing the genetic risk to develop age-related diseases, such as cardiovascular, autoimmune and neurodegenerative disorders, frailty, and cancer. Several GWAS have been performed on longevity, and although the replication rate across multiple study is generally low, multiple genomic regions across the genome have been linked to longevity. These genetic factors, along with environmental variables, model the relationship between longevity and age-related diseases, defining different aging trajectories. On the one end, protective elements (both genetic- and environmental-based) may shift towards older ages the onset of age-related diseases and compress disability periods, increasing the odds of successful aging. On the other, genetic predisposition to diseases and environmental risk-factors may accelerate the onset of age-related diseases, increasing late-life disability and shorthening lifespan.

comes that the more individuals are compared, the higher the chance to find significant associations.[16] However, an alternative strategy to increase statistical power may be to compare individuals that carry extreme phenotypes, as the effect-size should be maximal when comparing individuals that represent the extreme ends of a disease spectrum.[48, 49, 50] For AD and other age-related diseases, extreme cases may be defined by sporadic AD cases (*i.e.* cases with no familial AD) with a relatively early age at disease-onset. Extreme controls are represented by individuals who reach extreme ages without the disease. [48, 49, 50] Alternatively, instead of testing each variant independently, which is typically done in GWAS, one could test the combined effect of multiple variants that are associated with a certain disease. One of the most commonly used methods to test the combined effect of multiple common variants is the construction of polygenic risk scores (PRS).[51, 52] A PRS is a weighted score that quantifies the individual risk to a certain phenotype and therefore can be used to stratify patients according to their genetic risk for a given trait or to identify individuals at the highest genetic risk. Normally, PRS are constructed using genetic variants that are identified through GWAS, and under the assumption that genetic variants do not change over time (*i.e.* the effect of a genetic variant does not change at increasing ages), the PRS represents a powerful diagnostic and prognostic tool.[52]

Another drawback of GWAS relates to the interpretation of significant associations. The large majority of variants that are tested in GWAS are non-coding variants, for which the downstream effects on gene and protein function are unknown.[16] A trivial procedure is to associate the variant with the closest gene, assuming a linear organization of the DNA, which underestimates the complexities of our genome. Given the amount of data that is currently generated, multiple sources of variant annotation should be taken into consideration, such as expression-quantitative-trait-loci (eQTLs, *i.e.* associations between genetic variants and RNA expression), chromatin structure, or structural variations. Finally, to better understand how genetic factors affect different traits, it can be informative to explore the extent of association of a genomic region on different phenotypes. Altogether, it may be of interest to explore the added value of using extreme phenotypes in a case-control genetic analysis of AD, and to provide an innovative framework to perform gene-set enrichment analysis from set of SNPs.

1.6 Aim of this thesis and outline

The overall objective of this thesis is to investigate the genetic factors underlying extreme human longevity and the escape of Alzheimer's disease, for which we explore the genetic architecture of the cognitively healthy centenarians from the 100-plus Study (Figure 1.3). The thesis is subdivided into two sections: firstly, we focus on the comparison of the cognitively healthy agers with young AD cases and population controls in context of Alzheimer's disease and human longevity (chapter 2, chapter 3, chapter 4, and chapter 5). Secondly, we focus on the collaborative efforts in which our cohort participated in terms of large GWAS of AD and longevity, and is accompanied by the development of tools to integrate, visualize and analyze results from GWAS (chapter 6, chapter 7, and chapter 8).
We first explore the added value of analyzing extreme phenotypes in the genetic research of AD by comparing extreme controls, *i.e.* cognitively healthy centenarians, and extreme AD cases, *i.e.* relatively young AD cases, in a case-control study of AD (chapter 2). We report that cognitively healthy centenarians have a lower frequency of genetic variants associated with increased AD risk compared to the general population, and a higher frequency of protective variants. This led to a 2-fold enrichment in the variant effect-size when comparing AD cases with cognitively healthy agers, showing that the use of extreme phenotypes in genetic studies of complex traits is profitable.

We then investigate the molecular pathways that are known to play a role in AD pathogenesis and their association with resilience against AD, by combining the effect of multiple variants into polygenic risk scores (PRSs) and pathway-specific PRSs (chapter 3). We report that cognitively healthy centenarians have the lowest PRS and pathway-specific PRS for all major AD-associated pathways. Moreover, while the risk of AD was significantly associated with a higher pathway-specific PRS of all pathways, only the immune system response and endocytosis pathways significantly influenced the resilience against AD, even after excluding *APOE* variants.
In chapter 4, we challenge to disentangle the effect on healthy aging from the effect on AD risk of genetic variants that are associated with AD. Under the hypothesis that genetic variants increasing the risk of AD should negatively affect longevity, we found that most alleles that increase the risk of AD negatively influence healthy longevity, with the effect on AD that explained, for the majority of variants, the negative effect on healthy longevity. However, a subset of variants preferentially involved in immune-related processes

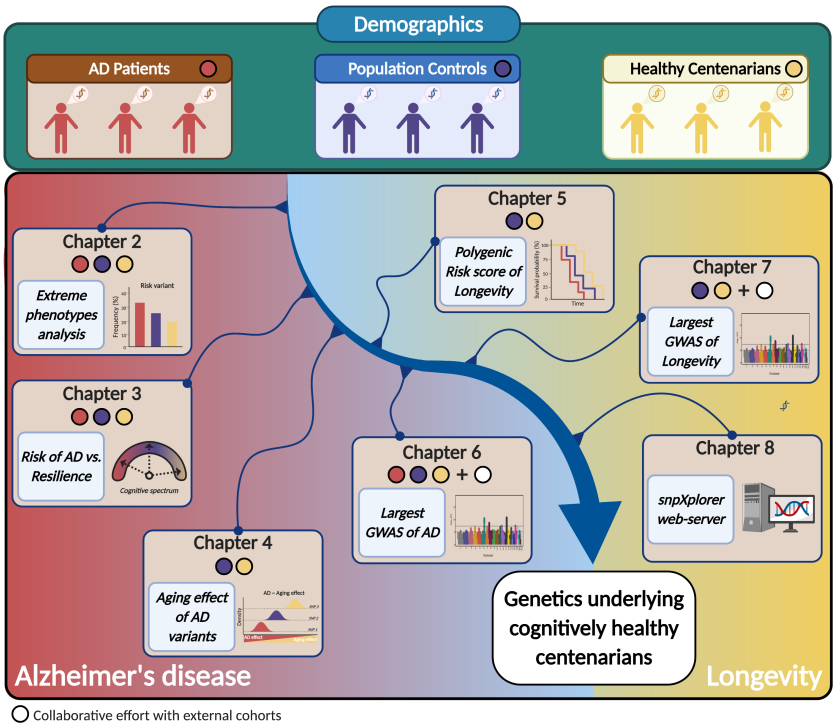


Figure 1.3: **Roadmap to understand the Genetics of Cognitively Healthy Centenarians.** All chapters in this thesis present analyses that are based on three distinct population: (*i*), a sample of Alzheimer's disease (AD) patients from the Amsterdam Dementia Cohort and other cohorts, (*ii*) a sample of healthy individuals from the different studies, and (*iii*) a sample of cognitively healthy centenarians from the 100-plus Study cohort. We will first focus on genetic variants associated with AD (chapter 2, chapter 3, chapter 4 and chapter 6), and then we will focus on the genetics of longevity (chapter 5, chapter 7 and chapter 8).

seemed to affect more strongly longevity than AD, suggesting a beneficial effect not only against AD, but also against other age-related diseases, or a general neuroprotective effect.
Then (chapter 5), we focus on human longevity and attempt to construct a polygenic risk score that associates with cognitively healthy aging and survival. Using the results from a study on parental longevity, we show that a polygenic risk score of 330 variants was significantly associated with becoming a cognitively healthy centenarian. Furthermore, this PRS significantly predicted survival in an independent cohort and was functionally enriched for biological pathways resembling the hallmarks of longevity, such as slow cell differentiation and replacement, and oxidative stress.

In the second part of the thesis, we present the contribution of the cognitively healthy centenarians from the 100-plus Study to large, collaborative GWAS of AD (chapter 6) and longevity (chapter 7). In chapter 6, we combined clinical studies and by-proxy studies in the largest GWAS of AD (at the time of publication), leading to the discovery of six additional genetic variants associated with AD. In addition to a better understanding of the genetic landscape of AD, our findings enforced the role of β-amyloid processing and immune response as central biological pathways in AD pathogenesis. Furthermore, we showed the applicability and predictability of the PRS for stratifying patients based on their genetic background and identifying those at the highest risk for the disease.
In chapter 7, we collaborated on, to date, the largest GWAS of longevity. We introduced an unbiased method to identify cases (*i.e.* long-lived individuals) and controls on country-based survival percentiles. In addition to *APOE* variants, we propose an additional variant near the *GPR78* gene to affect longevity, and through genetic correlation and gene expression analyses, we showed overlap between the genetics of diseases and the genetics of longevity.
Finally, (chapter 8) we developed *snpXplorer*, a tool that is freely available to the scientific community to explore summary statistics of genetic studies, compare levels of association between different traits, and functionally annotate sets of genetic variants. This tool may be useful to explore the extent of overlap between traits, which has applications in the diagnostic field.
The thesis ends with a summary and discussion.

References

[1] J. Oeppen. "DEMOGRAPHY: Enhanced: Broken Limits to Life Expectancy". In: *Science* 296.5570 (May 2002), pp. 1029–1031. ISSN: 00368075, 10959203. DOI: 10.1126/science.1069675.

[2] Department of Economic {and} Social Affair 2019. United Nations. *Retrieved from Profiles of Ageing 2019*. 2019.

[3] Linda Partridge, Joris Deelen, and P. Eline Slagboom. "Facing up to the global challenges of ageing". In: *Nature* 561.7721 (Sept. 2018), pp. 45–56. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0457-8.

[4] Emma Nichols et al. "Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016". In: *The Lancet Neurology* 18.1 (Jan. 2019), pp. 88–106. ISSN: 14744422. DOI: 10.1016/S1474-4422(18)30403-4.

[5] "2012 Alzheimer's disease facts and figures". In: *Alzheimer's & Dementia* 8.2 (Mar. 2012), pp. 131–168. ISSN: 15525260. DOI: 10.1016/j.jalz.2012.02.001.

[6] Henne Holstege et al. "The 100-plus Study of Dutch cognitively healthy centenarians: rationale, design and cohort description". In: (Apr. 2018). DOI: 10.1101/295287.

[7] Thomas Perls. "Dementia-free centenarians". In: *Experimental Gerontology* 39.11 (Nov. 2004), pp. 1587–1593. ISSN: 05315565. DOI: 10.1016/j.exger.2004.08.015.

[8] Bengt Winblad et al. "Defeating Alzheimer's disease and other dementias: a priority for European science and society". In: *The Lancet Neurology* 15.5 (Apr. 2016), pp. 455–532. ISSN: 14744422. DOI: 10.1016/S1474-4422(16)00062-4.

[9] María M. Corrada et al. "Dementia incidence continues to increase with age in the oldest old: The 90+ study". In: *Annals of Neurology* 67.1 (Jan. 2010), pp. 114–121. ISSN: 03645134, 15318249. DOI: 10.1002/ana.21915.

[10] Margaret Gatz et al. "Role of genes and environments for explaining Alzheimer disease". In: *Archives of General Psychiatry* 63.2 (Feb. 2006), pp. 168–174. ISSN: 0003-990X. DOI: 10.1001/archpsyc.63.2.168.

[11] A. M. Saunders et al. "Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease". In: *Neurology* 43.8 (Aug. 1993), pp. 1467–1472. ISSN: 0028-3878.

[12] W. J. Strittmatter et al. "Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease". In: *Proceedings of the National Academy of Sciences of the United States of America* 90.5 (Mar. 1993), pp. 1977–1981. ISSN: 0027-8424.

[13] E. H. Corder et al. "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families". In: *Science (New York, N.Y.)* 261.5123 (Aug. 1993), pp. 921–923. ISSN: 0036-8075.

[14] E. H. Corder et al. "Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease". In: *Nature Genetics* 7.2 (June 1994), pp. 180–184. ISSN: 1061-4036. DOI: 10.1038/ng0694-180.

[15] William S. Bush and Jason H. Moore. "Chapter 11: Genome-wide association studies". In: *PLoS computational biology* 8.12 (2012), e1002822. ISSN:

1553-7358. DOI: 10.1371/journal.pcbi.1002822.

[16] Vivian Tam et al. "Benefits and limitations of genome-wide association studies". In: *Nature Reviews Genetics* 20.8 (Aug. 2019), pp. 467–484. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-019-0127-1.

[17] Sayantan Das et al. "Next-generation genotype imputation service and methods". In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1284–1287. ISSN: 1546-1718. DOI: 10.1038/ng.3656.

[18] Shane McCarthy et al. "A reference panel of 64,976 haplotypes for genotype imputation". In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1279–1283. ISSN: 1546-1718. DOI: 10.1038/ng.3643.

[19] Jean-Charles Lambert et al. "Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease". In: *Nature Genetics* 41.10 (Oct. 2009), pp. 1094–1099. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.439.

[20] J. C. Lambert et al. "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease". In: *Nature Genetics* 45.12 (Dec. 2013), pp. 1452–1458. ISSN: 1546-1718. DOI: 10.1038/ng.2802.

[21] Denise Harold et al. "Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease". In: *Nature Genetics* 41.10 (Oct. 2009), pp. 1088–1093. ISSN: 1546-1718. DOI: 10.1038/ng.440.

[22] Sudha Seshadri et al. "Genome-wide analysis of genetic loci associated with Alzheimer disease". In: *JAMA* 303.18 (May 2010), pp. 1832–1840. ISSN: 1538-3598. DOI: 10.1001/jama.2010.574.

[23] Rebecca Sims et al. "Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease". In: *Nature Genetics* 49.9 (Sept. 2017), pp. 1373–1384. ISSN: 1546-1718. DOI: 10.1038/ng.3916.

[24] Rita Guerreiro et al. "TREM2 variants in Alzheimer's disease". In: *The New England Journal of Medicine* 368.2 (Jan. 2013), pp. 117–127. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1211851.

[25] Thorlakur Jonsson et al. "Variant of TREM2 associated with the risk of Alzheimer's disease". In: *The New England Journal of Medicine* 368.2 (Jan. 2013), pp. 107–116. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1211103.

[26] Paul Hollingworth et al. "Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease". In: *Nature Genetics* 43.5 (May 2011), pp. 429–435. ISSN: 1546-1718. DOI: 10.1038/ng.803.

[27] Adam C. Naj et al. "Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease". In: *Nature Genetics* 43.5 (May 2011), pp. 436–441. ISSN: 1546-1718. DOI: 10.1038/ng.801.

[28] G. Jun et al. "A novel Alzheimer disease locus located near the gene encoding tau protein". In: *Molecular Psychiatry* 21.1 (Jan. 2016), pp. 108–117. ISSN: 1476-5578. DOI: 10.1038/mp.2015.23.

[29] Stacy Steinberg et al. "Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease". In: *Nature Genetics* 47.5 (May 2015),

pp. 445–447. ISSN: 1546-1718. DOI: 10.1038/ng.3246.

[30] Céline Bellenguez et al. *New insights on the genetic etiology of Alzheimer's and related dementia*. en. preprint. Neurology, Oct. 2020. DOI: 10.1101/2020.10.01.20200659.

[31] J. Hardy. "The Amyloid Hypothesis of Alzheimer's Disease: Progress and Problems on the Road to Therapeutics". In: *Science* 297.5580 (July 19, 2002), pp. 353–356. ISSN: 00368075, 10959203. DOI: 10.1126/science.1072994.

[32] Caroline Van Cauwenberghe, Christine Van Broeckhoven, and Kristel Sleegers. "The genetic landscape of Alzheimer disease: clinical implications and perspectives". In: *Genetics in Medicine* 18.5 (May 2016), pp. 421–430. ISSN: 1098-3600, 1530-0366. DOI: 10.1038/gim.2015.117.

[33] R. M. Ransohoff. "How neuroinflammation contributes to neurodegeneration". In: *Science* 353.6301 (Aug. 2016), pp. 777–783. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aag2590.

[34] Michael T Heneka et al. "Neuroinflammation in Alzheimer's disease". In: *The Lancet Neurology* 14.4 (Apr. 2015), pp. 388–405. ISSN: 14744422. DOI: 10.1016/S1474-4422(15)70016-5.

[35] Nuria Garatachea et al. "ApoE gene and exceptional longevity: Insights from three independent cohorts". In: *Experimental Gerontology* 53 (May 2014), pp. 16–23. ISSN: 05315565. DOI: 10.1016/j.exger.2014.02.004.

[36] Hui Shi et al. "Genetic variants influencing human aging from late-onset Alzheimer's disease (LOAD) genome-wide association studies (GWAS)". In: *Neurobiology of Aging* 33.8 (Aug. 2012), 1849.e5–1849.e18. ISSN: 01974580. DOI: 10.1016/j.neurobiolaging.2012.02.014.

[37] J. Graham Ruby et al. "Estimates of the Heritability of Human Longevity Are Substantially Inflated due to Assortative Mating". In: *Genetics* 210.3 (Nov. 2018), pp. 1109–1124. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.118.301613.

[38] Joanna Kaplanis et al. "Quantitative analysis of population-scale family trees with millions of relatives". In: *Science* 360.6385 (Apr. 13, 2018), pp. 171–175. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aam9309.

[39] Paola Sebastiani and Thomas T. Perls. "The genetics of extreme longevity: lessons from the new England centenarian study". In: *Frontiers in Genetics* 3 (2012), p. 277. ISSN: 1664-8021. DOI: 10.3389/fgene.2012.00277.

[40] Marian Beekman et al. "Genome-wide association study (GWAS)-identified disease risk alleles do not compromise human longevity". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.42 (Oct. 2010), pp. 18046–18049. ISSN: 1091-6490. DOI: 10.1073/pnas.1003540107.

[41] Joris Deelen et al. "Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age". In: *Human Molecular Genetics* 23.16 (Aug. 2014), pp. 4420–4432. ISSN: 1460-2083. DOI: 10.1093/hmg/ddu139.

[42] Kristen Fortney et al. "Genome-Wide Scan Informed by Age-Related Disease Identifies Loci for Exceptional Human Longevity". In: *PLOS Genetics* 11.12 (Dec. 2015). Ed. by Hao Li,

e1005728. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1005728.

[43] Peter K. Joshi et al. "Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity". In: *Nature Communications* 8.1 (Dec. 2017). ISSN: 2041-1723. DOI: 10.1038/s41467-017-00934-5.

[44] Paola Sebastiani et al. "Four Genome-Wide Association Studies Identify New Extreme Longevity Variants". In: *The Journals of Gerontology: Series A* 72.11 (Oct. 2017), pp. 1453–1464. ISSN: 1079-5006, 1758-535X. DOI: 10.1093/gerona/glx027.

[45] Yi Zeng et al. "Novel loci and pathways significantly associated with longevity". In: *Scientific Reports* 6.1 (Aug. 2016). ISSN: 2045-2322. DOI: 10.1038/srep21243.

[46] Paul RHJ Timmers et al. "Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances". In: *eLife* 8 (Jan. 2019). ISSN: 2050-084X. DOI: 10.7554/eLife.39856.

[47] David Melzer, Luke C. Pilling, and Luigi Ferrucci. "The genetics of human ageing". In: *Nature Reviews Genetics* (Nov. 2019). ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-019-0183-6.

[48] Cristina Giuliani et al. "Centenarians as extreme phenotypes: An ecological perspective to get insight into the relationship between the genetics of longevity and age-associated diseases". In: *Mechanisms of Ageing and Development* 165 (July 2017), pp. 195–201. ISSN: 00476374. DOI: 10.1016/j.mad.2017.02.007.

[49] Paolo Garagnani et al. "Centenarians as super-controls to assess the biological relevance of genetic risk factors for common age-related diseases: a proof of principle on type 2 diabetes". In: *Aging* 5.5 (May 2013), pp. 373–385. ISSN: 1945-4589. DOI: 10.18632/aging.100562.

[50] Dalin Li et al. "Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies". In: *Genetic Epidemiology* 35.8 (Dec. 2011), pp. 790–799. ISSN: 1098-2272. DOI: 10.1002/gepi.20628.

[51] Frank Dudbridge. "Power and Predictive Accuracy of Polygenic Risk Scores". In: *PLoS Genetics* 9.3 (Mar. 2013). Ed. by Naomi R. Wray, e1003348. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003348.

[52] Ali Torkamani, Nathan E. Wineinger, and Eric J. Topol. "The personal and clinical utility of polygenic risk scores". In: *Nature Reviews Genetics* 19.9 (Sept. 2018), pp. 581–590. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-018-0018-x.

2. Extreme phenotypes

Centenarian controls increase variant effect sizes by an average twofold in an extreme case–extreme control analysis of Alzheimer's disease

Niccolo' Tesi *, Sven J. van der Lee *, Marc Hulsman, Iris E. Jansen, Najada Stringa, Natasja M. van Schoor, Martijn Huisman, Philip Scheltens, Marcel J.T. Reinders, Wiesje M. van der Flier and Henne Holstege

* Authors contributed equally

This chapter was published in *European Journal of Human Genetics*
https://doi.org/10.1038/s41431-018-0273-5

Abstract

The detection of genetic loci associated with Alzheimer's disease (AD) requires large numbers of cases and controls because variant effect sizes are mostly small. We hypothesized that variant effect sizes should increase when individuals who represent the extreme ends of a disease spectrum are considered, as their genomes are assumed to be maximally enriched or depleted with disease-associated genetic variants. We used 1,073 extensively phenotyped AD cases with relatively young age at onset as extreme cases (66.3 ± 7.9 years), 1,664 age-matched controls (66.0 ± 6.5 years) and 255 cognitively healthy centenarians as extreme controls (101.4 ± 1.3 years). We estimated the effect size of 29 variants that were previously associated with AD in genome-wide association studies. Comparing extreme AD cases with centenarian controls increased the variant effect size relative to published effect sizes by on average 1.90-fold (SE = 0.29, p = 0.0009). The effect size increase was largest for the rare high-impact *TREM2*(*R74H*) variant (6.5-fold), and significant for variants in/near *ECHDC3* (4.6-fold), *SLC24A4 – RIN3* (4.5-fold), *NME8* (3.8-fold), *PLCG2* (3.3-fold), *APOE – ϵ2* (2.2-fold), and *APOE – ϵ4* (twofold). Comparing extreme phenotypes enabled us to replicate the AD association for 10 variants ($p < 0.05$) in relatively small samples. The increase in effect sizes depended mainly on using centenarians as extreme controls: the average variant effect size was not increased in a comparison of extreme AD cases and age-matched controls (0.94-fold, p=0.68), suggesting that on average the tested genetic variants did not explain the extremity of the AD cases. Concluding, using centenarians as extreme controls in AD case–control studies boosts the variant effect size by on average twofold, allowing the replication of disease-association in relatively small samples.

2.1 Introduction

Alzheimer's disease (AD) is often characterized by a slow but progressive loss of cognitive functions, leading to loss of autonomy.[1] AD is rare at the age of 65 years, but its incidence increases exponentially to 40% at the age of 100 years.[2] It is currently the most prevalent cause of death at old age and one of the major health threats of the 21st century.[1] Better understanding of the etiological factors that determine AD is warranted as no treatment is currently available. Heritability plays an important role, as genetic factors are estimated to deter- mine 60–80% of the risk of AD. [3] About 30% of the genetic risk is attributable to the $\epsilon 4$ allele of *APOE* gene, and large collaborative efforts have identified over two dozen additional genetic loci that are associated with a slight modification of the risk of AD.[4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17] The design of these association studies relies on the comparison of very large numbers of cases with age-matched controls, such that detected associations can be attributed specifically to the disease.[18] However, given the prevalence of AD in the aging population, it is likely that a significant fraction of the controls will develop the disease at a later age. Therefore, as the AD risk for future cases likely involves the same genetic variants, using age-matched controls may quench variant association signals. This may, in part, explain the mostly small variant effect sizes associated with common variants. Also, GWAS studies mostly compare common genetic variants that are widely propagated in the population; as a consequence, these have mostly small effects on AD risk.[19] Rare genetic variants often have larger effect sizes than com- mon variants, but as there are fewer carriers available in the population, the requirement for large sample sizes stands.[20] The power of genetic analyses is determined by the variant frequency, the effect size of the variant, the sample size, and significance threshold set to be obtained.[21] Therefore, instead of increasing sample sizes of genetic studies to detect novel disease-associated genetic loci, an alternative strategy is to increase variant effect sizes by sampling individuals with extreme phenotypes.[20, 22, 23] For AD and other age-related diseases, extreme cases may be defined by having a relatively early age at disease onset, and having the phenotypic features characteristic for the disease, as defined by diagnostic assessment. Extreme controls are represented by individuals who reach extreme ages without the disease.[22, 24, 25] Indeed, in a case–control study of type 2 diabetes, the effect sizes for variants that were previously associated with the disease were increased when using centenarians as extreme controls.[24] The effect of using extreme phenotypes in other

age-related diseases has not been studied. Here, we explored the potential of using extreme phenotypes for genetic studies of AD by investigating the change in effect size of known AD-associated variants. Furthermore, using an age- and population-matched reference group, we investigated the contribution of each extreme phenotype.

Table 2.1: **Population characteristics**

	Extreme AD (EA)	Centenarian controls (EC)	Normal controls (NC)
Individuals	1,073	255	1,664
Females (%)	564 (52.6%)	191 (74.9%)	893 (53.7%)
Age (SD)[a]	66.4 (7.8)	101.4 (1.3)	66.0 (6.5)
APOE $\epsilon 4$ (%)	981 (42.7)	44 (8.6)	533 (16.0)
APOE $\epsilon 2$ (%)	76 (3.5)	78 (15.3)	304 (9.1)

[a]Age at onset for extreme Alzheimer's disease cases, age at study inclusion for extreme controls and normal controls; SD, standard deviation; *ApoE*, Apolipoprotein E allele count for $\epsilon 4$ and $\epsilon 2$, respectively. Reference to the cohorts reported in this table are: [18, 19, 20]

2.2 Methods

2.2.1 Cohort description

As extreme AD cases group (denoted by *EA*), we used 1,149 AD cases from the Amsterdam Dementia Cohort (ADC). The ADC comprises patients who visit the memory clinic of the VU University Medical Center, The Netherlands.[21, 18] This cohort of AD patients is extensively characterized and comprises 503 early-onset cases (denoted by *eEA*) with an age at onset <65 years, and 646 late-onset cases (denoted by *lEA*). Of the 503 early-onset cases, 255 had an age at onset <60 years (*i.e.*, young early onset, denoted by *yEA*). The diagnosis of probable AD was based on the clinical criteria formulated by the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) and based on National Institute of Aging-Alzheimer association (NIA-AA).[22, 23] At baseline, all subjects underwent a standard clinical diagnostic assessment including neurological examination and standard blood tests. In addition, all subjects underwent magnetic resonance imaging, an electroencephalogram, and cerebrospinal fluid (CSF) was analyzed for most patients.[24] Clinical diagnosis is made in consensus-based, multidisciplinary meetings. Together, this elaborate diagnostic procedure reduces the chance of misdiagnosis. The extensive phenotyping in combination with the early disease onset generates an AD cohort that can be regarded *extreme*. As extreme control group (denoted by *EC*), we used 268 self-reported cognitively healthy centenarians from the 100-plus Study cohort.[20] This study includes Dutch-speaking individuals who (*i*) can provide official evidence for being aged 100 years or older, (*ii*) self-report to be cognitively healthy, which is confirmed by a proxy, (*iii*) consent to donation of a blood sample, (*iv*) consent

to (at least) two home visits from a researcher, and (*v*) consent to undergo an interview and neuropsychological test battery. As *normal controls* (denoted by *NC*) we used 1,717 middle-aged (55–85 year-old) individuals from a representative sample of Dutch individuals from the Longitudinal Aging Study Amsterdam (LASA) cohort.[19, 25] LASA is an ongoing longitudinal study of older adults initiated in 1991, with the main objective to determine predictors and consequences of aging. The Medical Ethics Committee of the VU University Medical Center (METC) approved the ADC cohort, the LASA study and the 100-plus Study. All participants and/or their legal guardians gave written informed consent for participation in clinical and genetic studies.

2.2.2 Genotyping and imputation methods

We selected 29 single-nucleotide variants for which evidence for a genome-wide significant association with AD was found in previous studies (Table S1, Table S2).[4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17] Genetic variants were determined by standard genotyping or imputation methods. In brief, we genotyped all individuals using the Illumina Global Screening Array (GSAsharedCUSTOM_20018389_A2) and applied established quality control methods.[26] We used high-quality genotyping in all individuals (individual call rate > 98%, variant call rate > 98%), individuals with sex mismatches were excluded and Hardy-Weinberg equilibrium-departure was considered significant at $p < 1\text{x}10^{-6}$. Genotypes were prepared for imputation using provided scripts (HRC-1000G-check-bim.pl).[27] This script compares variant ID, strand and allele frequencies to the haplotype reference panel (HRC v1.1, April 2016).[26] Finally, all autosomal variants were submitted to the Michigan imputation server (`https://imputationserver.sph.umich.edu`).[26] The server uses SHAPEIT2 (v2.r790) to phase data and imputation to the reference panel (v1.1) was performed with Minimac3.[26, 28] A total of 1,149 extreme AD cases, 1,717 normal controls and 268 extreme (centenarian) controls passed quality control. Prior to analysis, we excluded individuals of non-European ancestry (N_{EA} = 67, based on 1000Genomes [29] clustering) and individuals with a family relation (N_{EA}=9, N_{EC}=13, N_{NC}=53, identity-by-descent ≥0.3), [30] leaving 1,073 extreme AD cases (N_{eEA}=464 and N_{lEA}=609), 1,664 normal controls and 255 centenarian controls for the analysis.

Table 2.2: **Association statistics of the 29 tested AD-associated variants**

Chr	Position	Rs ID	Gene	A1	$\log OR_I^k$(SE)	$\log OR_{EA-EC}^k$(SE)	P_{EA-EC}	E_{EA-EC}^k(95% CI, p)	AF_{EA}	AF_{NC}	AF_{EC}
6	41,129,252	rs75932628	*TREM2 (R47H)*	T	0.89 (0.09)	5.75 (5.13)	0.26	6.46 (-4.83 - 18.46, 0.35)	0.003	0.001	0.000
10	11,720,308	rs7920721	*ECHDC3*	G	0.07 (0.01)	0.31 (0.10)	0.0029*	4.56 (1.55 - 8.93, 0.018)	0.430	0.389	0.357
14	92,926,952	rs10498633	*SLC24A4-RIN3*	T	-0.09 (0.01)	-0.42 (0.11)	0.0001*	4.50 (2.08 - 7.93, 0.0028)	0.206	0.236	0.292
7	37,841,534	rs2718058	*NME8*	G	-0.08 (0.01)	-0.29 (0.10)	0.0037*	3.80 (1.17 - 7.28, 0.033)	0.360	0.367	0.433
16	81,942,028	rs72824905	*PLCG2*	G	-0.39 (0.06)	-1.27 (0.40)	0.0014*	3.28 (1.26 - 5.98, 0.028)	0.008	0.012	0.025
6	32,578,530	rs9271192	*HLA-DRB1*	A	-0.11 (0.01)	-0.35 (0.16)	0.031*	3.20 (0.35 - 6.65, 0.13)	0.712	0.727	0.780
7	100,004,446	rs1476679	*ZCWPW1*	T	0.09 (0.01)	0.26 (0.11)	0.013*	2.97 (0.60 - 6.10, 0.10)	0.703	0.674	0.649
19	1,063,443	rs4147929	*ABCA7 (A>G)*	G	-0.14 (0.02)	-0.32 (0.14)	0.021*	2.26 (0.30 - 4.42, 0.22)	0.809	0.834	0.855
19	45,412,079	rs7412	*APOE ($\epsilon 2$)*	T	-0.79 (0.03)	-1.76 (0.18)	$3.16x10^{-21}$*	2.24 (1.75 - 2.77, $1.4x10^{-7}$)	0.033	0.091	0.149
4	11,711,232	rs13113697	*HS3ST1*	G	-0.07 (0.01)	-0.14 (0.12)	0.24	2.06 (-1.49 - 6.13, 0.54)	0.265	0.268	0.247
17	47,297,297	rs616338	*ABI3*	C	-0.36 (0.05)	-0.74 (0.57)	0.19	2.06 (-0.99 - 5.59, 0.52)	0.017	0.009	0.006
6	47,487,762	rs10948363	*CD2AP*	G	0.10 (0.01)	0.19 (0.11)	0.088	2.00 (-0.34 - 4.60, 0.41)	0.284	0.272	0.245
19	45,411,941	rs429358	*APOE ($\epsilon 4$)*	C	1.05 (0.03)	2.08 (0.17)	$1.31x10^{-33}$*	1.99 (1.65 - 2.33, $1.5x10^{-9}$)	0.429	0.166	0.082
7	143,110,762	rs11771145	*EPHA1*	A	-0.10 (0.01)	-0.20 (0.10)	0.059	1.94 (-0.09 - 4.29, 0.37)	0.325	0.345	0.371
11	47,557,871	rs10838725	*CELF1*	C	0.08 (0.01)	0.14 (0.11)	0.205	1.78 (-0.95 - 5.11, 0.58)	0.328	0.314	0.302
8	27,195,121	rs28834970	*PTK2B*	C	0.10 (0.01)	0.18 (0.10)	0.089	1.76 (-0.23 - 4.09, 0.47)	0.395	0.376	0.353
11	59,923,508	rs983392	*MS4A6A*	G	-0.11 (0.01)	-0.17 (0.10)	0.094	1.56 (-0.20 - 3.61, 0.54)	0.397	0.403	0.439
11	121,435,587	rs11218343	*SORL1*	C	-0.26 (0.03)	-0.39 (0.25)	0.12	1.48 (-0.39 - 3.51, 0.62)	0.033	0.040	0.047
2	127,892,810	rs6733839	*BIN1*	T	0.20 (0.01)	0.25 (0.10)	0.011*	1.28 (0.31 - 2.29, 0.58)	0.456	0.413	0.390
11	85,867,875	rs10792832	PICALM	G	0.14 (0.01)	0.15 (0.10)	0.13	1.09 (-0.30 - 2.56, 0.91)	0.653	0.614	0.612
20	55,018,260	rs7274581	*CASS4*	C	-0.13 (0.02)	-0.14 (0.18)	0.44	1.06 (-1.83 - 4.07, 0.97)	0.075	0.088	0.084
6	41,129,207	rs143332484	*TREM2 (R62H)*	T	0.50 (0.07)	0.48 (0.48)	0.32	0.97 (-0.96 - 3.09, 0.98)	0.017	0.015	0.009
17	44,353,222	rs118172952	*KANSL1*	G	-0.14 (0.03)	-0.13 (0.14)	0.34	0.97 (-1.08 - 3.64, 0.96)	0.191	0.202	0.221
1	207,692,049	rs6656401	*CR1*	G	-0.17 (0.01)	-0.12 (0.12)	0.31	0.75 (-0.75 - 2.21, 0.74)	0.781	0.803	0.806
19	1,061,892	rs200538373	*ABCA7 (G>C)*	C	-0.65 (0.14)	-0.44 (0.80)	0.58	0.68 (-1.83 - 3.54, 0.79)	0.004	0.004	0.006
8	27,467,686	rs9331896	*CLU*	T	0.15 (0.01)	0.09 (0.10)	0.40	0.60 (-0.78 - 2.06, 0.58)	0.361	0.400	0.378
2	234,068,476	rs35349669	*INPP5D*	T	0.08 (0.01)	0.03 (0.10)	0.78	0.36 (-2.33 - 3.16, 0.62)	0.474	0.496	0.486
14	53,400,629	rs17125944	*FERMT2*	C	0.13 (0.02)	-0.11 (0.16)	0.50	-0.82 (-3.46 - 1.60, 0.13)	0.104	0.105	0.114
5	88,223,420	rs190982	*MEF2C*	A	0.08 (0.01)	-0.14 (0.10)	0.17	-1.86 (-5.01 - 0.77, 0.033)	0.408	0.406	0.372
		AVERAGE					1.90 ± 0.29, $p = 0.0009$				

Chr, chromosome; *Position*, chromosomal position in GRCh37; *Rs ID*, variant identifier; *Gene*, gene associated with the variant according to the original paper; *A1*, tested allele (alternative allele according to Haplotype Reference Consortium (HRC) panel); $\log OR_I^k$ *(SE)*, log(odds ratio) and relative standard error for variant k reported by study with largest sample size; $\log OR_{EA-EC}^k$*(SE)*,log(oddsratio) and relative standard error in extreme control association; P_{EA-EC},p-value of AD association of extreme AD cases *vs.* centenarian controls; E_{EA-EC}^k*(95% CI, p)*, change in effect size, 95% confidence intervals and *p-value* of difference when using extreme phenotypes relative to published effect sizes; AF_{EA}, tested allele frequency in AD extreme cases; AF_{NC}, tested allele frequency in normal controls; AF_{EC}, tested allele frequency in centenarian controls. Star*: significant at p<0.05

2.2.3 Statistical analysis

For each AD-associated variant, we explored the change in effect size (E) relative to reported effect sizes when (1) comparing extreme AD cases with extreme (centenarian) controls (*EA vs. EC*); (2) comparing extreme AD cases with normal controls (*EA vs. NC*); and (3) comparing normal AD cases with extreme (centenarian) controls (*NA vs. EC*). To calculate variant effect sizes, we used logistic regression models correcting for population stratification (principal components 1–6).[31, 32] We calculated odds ratios relative to the Haplotype Reference Consortium (HRC) alternative allele assuming additive genetic effects, and estimated 95% confidence intervals (CIs). We estimated the change in effect size relative to reported effect sizes (E) as follows:

$$E^k_{1-2} = \frac{\log OR^k_{1-2}}{\log OR^k_l} \tag{2.1}$$

where E^k_{1-2} indicates the effect size change for variant k in a comparison of cohort 1 and cohort 2, *e.g*, $E^{APOE\epsilon4}_{EA-EC}$ indicates the effect size change for the *APOE* $\epsilon4$ variant when extreme AD cases (*EA*) are compared with cognitive healthy centenarians (*EC*). The $\log OR^k_{1-2}$ denotes the effect size of variant k when comparing cohort 1 and cohort 2. The effect size of variant k reported in literature (Table S1) is denoted by $\log OR^k_l$. We estimated the added value of using extreme (centenarian) controls rather than normal age-matched controls in a case-control analysis. For this, we wanted to compute the change in effect size when comparing non-extreme AD cases with extreme controls (*NA vs. EC*). As we do not have direct access to *normal AD cases*, we estimated the effect size for the *NA-EC* comparison by summing (1) the effect size from the comparison of *normal AD cases* and *normal controls*, as reported in literature ($\log OR^k_l$), and (2) the effect size from the comparison of normal controls (*NC*) with extreme (centenarian) controls (*NC vs. EC*), *i.e.*, $\log OR^k_{NA-EC} = \log OR^k_l + \log OR^k_{NC-EC}$. The added value of using extreme controls in a case-control analysis then becomes:

$$E^k_{NA-EC} = \frac{\log OR^k_l + \log OR^k_{NC-EC}}{\log OR^k_k} \tag{2.2}$$

To assess whether age at disease onset had an impact on the change in effect size due to the extreme cases (E_{EA-NC}), we estimated the $\log OR^k_{eEA-NC}$ (early-onset extreme AD cases *vs.* normal controls), $\log OR^k_{lEA-NC}$ (late-onset extreme AD cases *vs.* normal controls) and the $\log OR^k_{yEA-NC}$ (younger

early-onset AD cases *vs.* normal controls), and their 95% CI. Then, we computed the probability that the effect size changes E^k_{eEA-NC} and E^k_{lEA-NC} differed using a two-samples z-test (two-tailed *p-value*).

2.2.4 Determining significance of change in effect size

For each variant, we estimated E^k_{1-2} and a 95% CI by sampling (S=10,000) from the $\log OR^k_{1-2}$ and $\log OR^k_l$ based on their respective standard errors. The probability of divergence between the distributions of the $\log OR^k_{1-2}$ and the $\log OR^k_l$ was determined using a two-sample z-test (two-tailed p-value). The probability of observing $E^k_{1-2} > 1$, *i.e*, an increased effect size for variant k, is considered to be a Bernoulli variable with p=0.5 (equal chance of having an increased/decreased effect). The number of variants that show an increase in effect (E^k_{1-2}) then follows a binomial distribution. The average change in effect size across all K=29 tested variants is calculated as follows:

$$\bar{E}_{1-2} = \frac{1}{K}\sum_{k}^{K} E^k_{1-2} \tag{2.3}$$

Confidence intervals and probability of divergence between $\bar{E}_{1-2}$ and previously reported effect sizes were estimated by sampling (S=10,000, two-tailed *p-value*). Quality control of genotype data, population stratification analysis, and relatedness analyses were performed with PLINK (v1.90b4.6), whereas association analysis, downstream analyses, and plots were performed with R (v3.3.2).[33, 34]

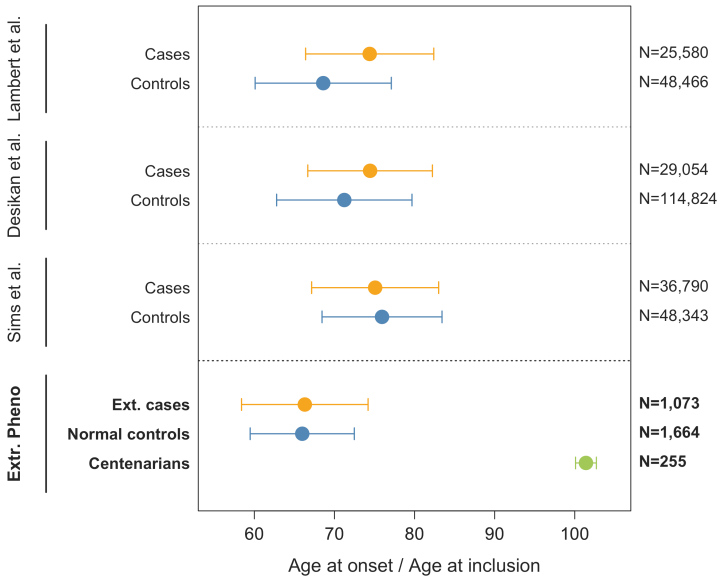


Figure 2.1: **Comparison of age at disease-onset and age at inclusion for cases and controls in previously reported case-control comparisons, and in our extreme phenotypes comparison.** Weighted mean and (combined) standard deviation of the age at onset for AD cases and age at inclusion for controls. As weights, we used the sample sizes of each GWA study. Note that previous case-control studies of AD included samples from multiple cohorts, sometimes overlapping across studies. References to the cohorts reported in this figure are: [7, 8, 13]

2.3 Results

After quality control of the genetic data, we included 1,073 extreme AD cases (with mean age at onset 66.4 ± 7.8 and 52.7% females), 1,664 normal (age-matched) controls (mean age at inclusion 66.0 ± 6.5, 53.7% females), and 255 cognitive healthy centenarians as extreme controls (mean age at inclusion 101.4 ± 1.3, 74.7% females) (Table 2.1). Within the extreme AD cases group, there were 464 early- onset cases (mean age at onset 59.1 ± 4.1, 54% females), and 609 late-onset cases (mean age at onset 72.1 ± 4.8, 51% females). The age at onset of the extreme AD cases was on average 8.2 years earlier compared with previous GWA studies; the age at disease onset was on average 15.4 years earlier in early-onset cases and 2.5 years earlier in late-onset cases, whereas the age at study inclusion of our centenarian controls was on average 29.5 years higher than for previously published

controls (Figure 2.1).

2.3.1 Effect of comparing extreme cases and centenarian controls

In a genetic comparison of extreme AD cases and centenarian controls (*EA–EC* comparison) the average effect size over all 29 genetic variants was 1.90-fold increased relative to the effect sizes reported in published studies ($\bar{E}_{EA-EC}$=1.90 ± 0.29; p=0.0009) (Figure 2.3). For 21 out of 29 variants, we observed an increased effect size (E^{k}_{EA-EC}>1),which is significantly more than expected by chance (p=0.012) (Figure 2.2 and Table 2.2). The increase in effect size ranged from 1.06 (variant near *CASS4*) to 6.46 (variant in *TREM2 [R47H]*) and was observed both in common variants (MAF>1%, n = 19) and rare variants (MAF<1%; *TREM2 [R47H]* and *ABI3*) (Table 2.2). For variants near or in the genes *TREM2 (R47H)*, *SLC24A4-RIN3*, and *ECHDC3*, the increase was more than fourfold compared with previously reported effect sizes. For nine variants the effect size increase was two- to fourfold (in or near the genes *NME8, PLCG2, HLA-DRB1, CD2AP, ZCWPW1, ABCA7 [A > G], APOE ϵ2, HS3ST1*, and *ABI3*, in order from high to low effect size increases). For nine variants the increase was between one- and twofold (in or near genes, *APOE ϵ4, EPHA1, CELF1, PTK2B, MS4A6A, SORL1, BIN1, PICALM*, and *CASS4*) (Figure 2.2). The effect sizes of six genetic variants were not increased in our extreme phenotype analysis compared with previously reported effect sizes ($\bar{E}_{EA-EC}$ between 0 and 1): in or near *TREM2 (R62H), KANSL1, CR1, ABCA7 (G > C), CLU*, and *INPP5D*. At last, the effect sizes of two variants were in the opposite direction compared to previously reported effects(E^{k}_{EA-EC}). Specifically, for the variant in *FERMT2* we found an inverted direction of effect size and a lower magnitude of effect as compared with previous studies (E^{FERMT2}_{EA-EC} between 0 and -1). For the variant near *MEF2C* we observed a larger effect size as compared with those previously published, but in the opposite direction (E^{MEF2C}_{EA-EC}<1).

Overall, for seven common variants (MAF> 1%), the effect size was significantly increased relatively to the previously reported effect sizes (Table 2.2), in or near genes *APOE ϵ2* (2.2-fold, p=1.4x10^{-7}), *APOE ϵ4* (2.0-fold, p=1.5x10^{-9}), *SLC24A4-RIN3* (4.5-fold, p=2.8x10^{-3}), *ECHDC3* (4.6-fold, p=0.018), *PLCG2* (3.3-fold p=0.028), *NME8* (3.9-fold, p=0.033), and *MEF2C* (-1.9-fold, p=0.033). Variants with significant effect size changes were also more likely to be associated with AD in a comparison of extreme cases and centenarians. The association with AD reached nominal significance (p<0.05) in 10 out of 21 variants with a changed effect size >1 (Table 2.2). Next to *APOE ϵ4*

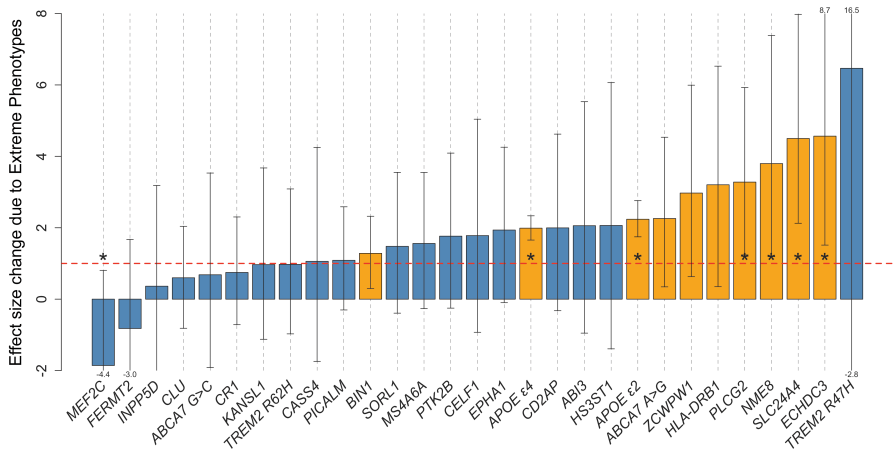


Figure 2.2: **Change in variant effect-size using extreme cases and centenarian controls relative to published effect-sizes, for 29 AD associated genetic variants.** Dashed red line at E^{k}_{EA-EC} = 1 indicates same effect- size as reported in literature. Orange bars indicate nominal statistical significance for the association with AD (p<0.05). Stars indicate significant changes of effect-size relative to previously reported effect sizes (p<0.05, two-sample z-test)

($\log OR^{APOE\epsilon4}_{EA-EC}$=2.1, SE=0.17, p=1.3x10^{-33}) and *APOE ϵ2* ($\log OR^{APOE\epsilon2}_{EA-EC}$=-1.8, p=3.2x10^{-21}), variants in or near these genes were significantly associated with AD: *SCL24A4-RIN3, PLCG2, ECHDC3, NME8, BIN1, ZCWPW1, ABCA7 (A > G)*, and *HLA-DRB1* (Table 2.2).

2.3.2 Effect of using extreme AD cases

The average effect size in a comparison of extreme AD cases with normal controls (*EA vs. NC*) did not significantly change relative to the previously reported effect sizes ($\bar{E}_{EA-EC}$=0.94 ± 0.12, p=0.68) (Figure 2.3). The effect size was significantly increased for *APOE ϵ4* variant (1.3-fold, p=1.4x10^{-5}), and nominally significant for *APOE ϵ2* (1.4-fold, p=0.017). For 14 individual variants, we observed an increased effect size, but this was not more than what could be expected by chance (p=0.5, Figure 2.4 and Table S3). We then separated AD cases into early-onset extreme AD cases (N_{eEA}=464, age at onset <65 years) and late-onset extreme AD cases (N_{lEA}=609), and estimated the change in effect sizes. Unexpectedly, the average effect size in the early-onset cases was lower relative to previously published effect sizes ($\bar{E}_{eEA-NC}$ was 0.86 ± 0.16, p=0.79), whereas for late-onset cases the effect size was

similar to published effect sizes ($\bar{E}_{lEA-NC}$ was 1.01 ± 0.14, p=0.46) (Figure 2.6 and Table S4). We found significant differences between the effect sizes in early-onset and late-onset AD cases ($\log OR^k_{eEA-NC}$ and $\log OR^k_{lEA-NC}$, respectively) for the variants in or near *APOE ϵ2* (-0.41 *vs.* -0.89; p=0.05), *ZCWPW1* (0.01 *vs.* 0.24: p=0.016) and *MS4A6A* (0.12 *vs.* -0.13; p=0.0079). When we extended the comparison with only the youngest early-onset AD cases (N_{yEA}=255, age at onset <60 years) and normal controls, the average effect size was still lower than previously published effect sizes ($\bar{E}_{yEA-NC}$ was 0.87 ± 0.20, p=0.74) (Table S4).

2.3.3 Effect of extreme controls

In a comparison of normal AD cases and extreme (centenarian) controls (*NA vs. EC*), the effect size was on average 1.88-fold higher relative to previously reported effect sizes ($\bar{E}_{NA-EC}$ = 1.88 ± 0.24, p=0.0001) (Figure 2.3 and Figure 2.5). This was almost identical to the average increase in effect size when we compared the extreme cases with centenarian controls ($\bar{E}_{EA-EC}$ = 1.90 ± 0.29; p=0.0009) (Figure 2.3). At the variant level, the change in effect sizes was similar in both analyses (Figure 2.7A). In fact, in a comparison of normal AD cases with extreme controls, we observed an increased effect size for 24/29 variants relative to published variant effect sizes (E^k_{NA-EC}> 1), which is more than expected by chance (p=0.00027) (Figure 2.5 and Table S3). As in the comparison of the extremes, we found a significant increase in effect size for variants in or near *APOE ϵ2* (1.7-fold, p<5x10^{-5}), *APOE ϵ4* (1.7-fold, p<5x10^{-5}), *NME8* (4.5-fold, p=0.0035), *SLC24A4-RIN3* (3.9-fold, p=0.0045) and *PLCG2* (2.9-fold, p=0.019). The main exception to this was the increased effect size of the rare *TREM2 (R47H)* variant (allele frequency = 0.001), which was increased more when using extreme AD cases than when using normal AD cases in a comparison with extreme controls (6.46-fold *vs.* 3.42-fold) (Figure 2.7A). For this rare variant we identified seven carriers in 1,073 extreme cases, and none in 255 centenarian controls. The effect size increase did not reach significance as CIs were large, which is according to expectations for very rare variants in small sample sizes. However, overall, the extreme controls contributed more to the effect size change than the extreme cases in a comparison of the extremes (Figure 2.7B).

2.4 Discussion

In this study, we found that the effect sizes of 29 variants previously identified in genetic case control analyses for AD were increased in a case–control analysis of extreme phenotypes. The use of extreme AD cases and cognitively healthy centenarians as extreme controls increased effect sizes for association with AD up to sixfold, relative to previously published effect sizes. On average, the use of extreme phenotypes almost doubled the variant effect size. Although changes in effect size were different per variant, the effect size increase was driven mainly by the centenarian controls. This profound increase enabled us to replicate the association with AD of 10 common variants in relatively small samples. In a comparison of AD cases (either normal or extreme) with centenarian controls, we observed significant effect size increases for variants in or near *PLCG2, NME8, ECHDC3, SLC24A4-RIN3, APOE $\epsilon 2$*, and *APOE $\epsilon 4$*. We also found a large effect size increase for the rare *TREM2 (R47H)* risk variant, which did not reach significance owing to variant rareness. This suggests that the tested variants or loci might (positively or negatively) contribute to the long-term preservation of cognitive health and/or to longevity in general. *PLCG2, NME8*, and *TREM2* are implicated in immunological processes,[8, 35] whereas *SLC24A4, ECHDC3,* and *APOE* are involved in lipid and cholesterol metabolism (Table S5).[17, 36, 37] Both these processes were previously associated with longevity,[38, 39] such that an overlapping etiology of maintained cognitive health and maintained overall health may contribute to the observed increase in effect size. However, with the exception of the *APOE* locus, these loci were thus far not associated with longevity in GWA studies.[40, 41, 42, 43] We speculate that the association might be dependent on the maintained cognitive health in the centenarians of the 100-plus Study cohort.[20] Alternatively, longevity studies may have been underpowered to detect the association of these loci with extreme survival. Future studies will have to establish the mechanism behind the association of these genes with preserved cognitive health.

Next to *APOE*, the *HLA-DRB1* locus has been associated with both AD [13] and longevity.[40] However, its most informative variants, rs9271192 for AD and rs34831921 for longevity, are not in linkage disequilibrium (r^2 = 0.04), suggesting that these are independent signals. Interestingly, the variants for which the effect size did not significantly increase when using extreme cases and centenarian controls are also involved in immunity (variants in/near *TREM2, CR1, ABCA7, CLU, INPP5D*, and *MEF2C*) and lipid/cholesterol metabolism (variants in/near *ABCA7* and *CLU*) (Table S5). We speculate

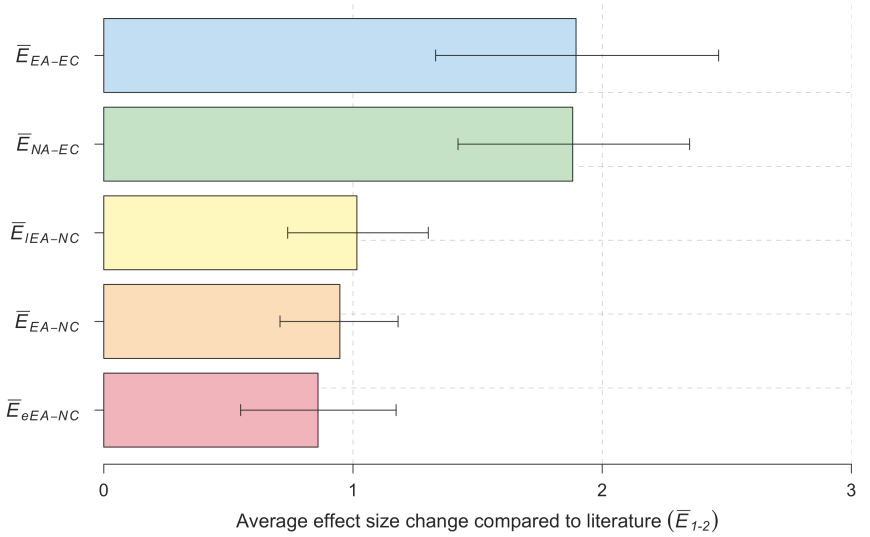


Figure 2.3: **Average increase in effect-size for the different comparisons.** Average increase in effect sizes for: Extreme AD cases (N_{EA} = 1,073), of which early onset cases (N_{eEA} = 464), late onset cases (N_{lEA} = 609); centenarian controls (N_{EC} = 255); normal controls (N_{NC} = 1,664). 95% confidence intervals were estimated by random sampling (S = 10,000)

that the variants with an increased effect size might influence changes in cognitive health during aging while variants with no increased effect size do not influence these processes. Using extreme cases did not increase the variant effect sizes relative to published effect sizes, even though most of the extreme cases were biomarker confirmed and their mean age at onset was 8.2 years younger than the mean age at onset in other studies.[7, 8, 13] The only exception to this was the (non-significant) effect size increase for the rare *TREM2 (R47H)* risk variant, which was driven in part by using extreme AD cases. This suggests that based on the tested genetic variants, the *phenotypically extreme* cases presented in this study are not genetically more extreme than cases presented in other studies. In fact, the variant effect sizes of early-onset AD cases were on average lower than the variant effect size of late-onset AD cases, and this persisted even when selecting only the youngest early-onset cases. One explanation for this observation may be that an early age at onset may be driven by rare, high-impact variants,[44] whereas the disease onset at later ages may depend to a greater extent on more common risk variants. Furthermore, we found significant differences at the variant level, between the effect sizes in early-onset and late-onset cases for common variants in/near *ZCWPW1* and *APOE $\epsilon 2$*, and also in -opposite directions- for the variant in *MS4A6A*. These results are a first indication that these variants may differentially influence age of disease onset, however, future experiments will have to confirm this finding. Our main finding is that, in a genetic case-control study of extreme phenotypes, the majority of the observed increase in effect size is attributable to the extreme controls, implicating that collecting cohorts of extreme controls is profitable. We note that the centenarians used in this study were selected for their preserved cognitive health, which might have further enlarged the effect size increase for genetic variants that were previously identified for their AD association. We acknowledge that using centenarians as controls in genetic studies of AD could result in the detection of variants associated with extreme longevity, such that newly detected AD-associations need to be verified in an age-matched AD case-control setting. Nevertheless, the effect sizes for all but two variants are in the same direction as previously reported, which suggests that the tested AD variants do not have significant pleiotropic activities that counteract their AD-related survival effects. Notably, the two variants with an opposite effect, in or near *MEF2C* and *FERMT2*, also did not associate with AD in our age-matched case-control analysis. This suggests that the AD association of these variants is not consistent across studies. This is in line with results from unpublished GWASs of AD in which AD-associations

of variants near the *MEF2C* and *FERMT2* genes were not replicated [45, 46] (p=0.053,[45] p=0.0003 for *MEF2C* [46] and p=1.6x10^{-5} for *FERMT2* [46] variant, with 5.0x10^{-8} being the genome-wide significance threshold). A strength of our study is that our cohorts of AD patients and controls, were not previously used in the discovery of any of the known AD-associated variants; we thus provide independent replication in a genetically homogeneous group of individuals, as they all came from one specific population (Dutch). Concluding, in our comparison of cases and controls with extreme phenotypes we found that on average, the effect of AD-related variants in genetic association studies almost doubled, whereas at the variant level effect sizes increased up to sixfold. The observed increment in effect size was driven by the centenarians as extreme controls, identifying centenarians as a valuable resource for genetic studies, with possible applications for other age-related diseases.

2.5 Acknowledgements

Research of the Alzheimer center Amsterdam is part of the neurodegeneration research program of Amsterdam Neuroscience (www.amsterdamresearch.org). The Alzheimer Center Amsterdam is supported by Stichting Alzheimer Nederland (WE09.2014-03) and Stiching VUmc fonds. The clinical database structure was developed with funding from Stichting Dioraphte (VSM 14 04 14 02). The Dutch case–control study is part of EADB (European Alzheimer DNA biobank) funded by JPcofundNL (ZonMW project number: 733051061). This work was in part carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. **Conflict of interest**: the authors declare no conflict of interest.

2.6 Full author list and affiliations

Niccolo' Tesi,[1,2,3] Sven J. van der Lee,[1,2] Marc Hulsman,[1,2,3] Iris E. Jansen,[1,4] Najada Stringa,[5] Natasja M. van Schoor,[5] Martijn Huisman,[5] Philip Scheltens,[1] Marcel J.T. Reinders,[3] Wiesje M. van der Flier,[1,5] and Henne Holstege [1,2,3]

[1] Alzheimer Centre, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[2] Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[3] Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

[4] Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, VU, Amsterdam, The Netherlands

[5] Department of Epidemiology and Data Sciences, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

2.7 Supplementary Figures

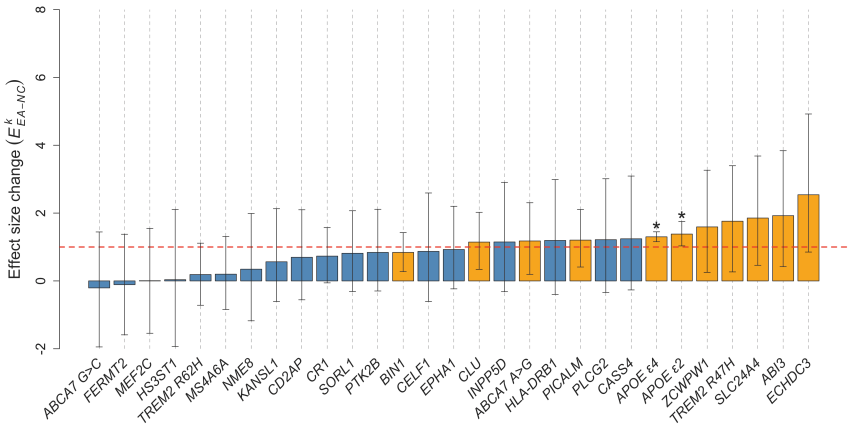


Figure 2.4: **Extreme AD cases vs. normal controls: E^k_{EA-NC}.** The effect-size change was significant for 4 variants (p<0.05, two-sample z-test; bars annotated with a star [*]). Orange bars indicate nominal statistical significance for the association with AD (p<0.05). Dashed red line ($E^k_{EA-NC} = 1$) indicates same effect-size as reported in literature.

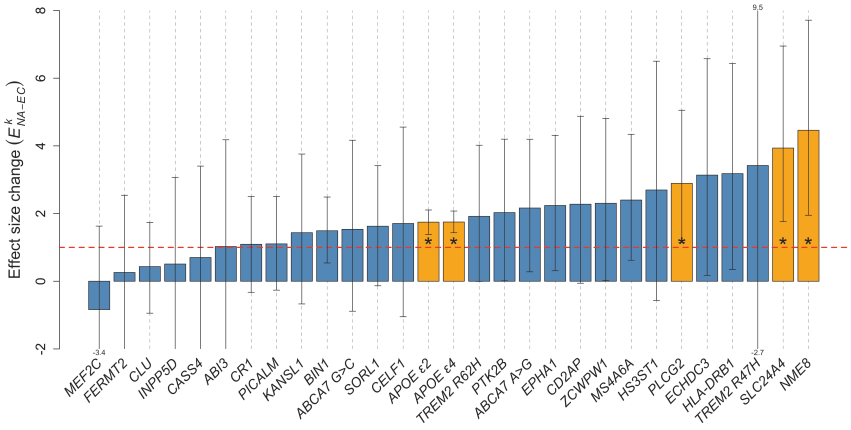


Figure 2.5: **Normal ADs vs. Extreme (centenarian) controls: E^k_{NA-EC}.** Effect-size change (E^k_{NA-EC}) was significant for 5 variants (p<0.05, two-sample z-test; bars annotated with a star [*]). Orange bars indicate nominal statistical significance for the association with AD (p<0.05). Dashed red line indicates same effect-size as reported in literature.

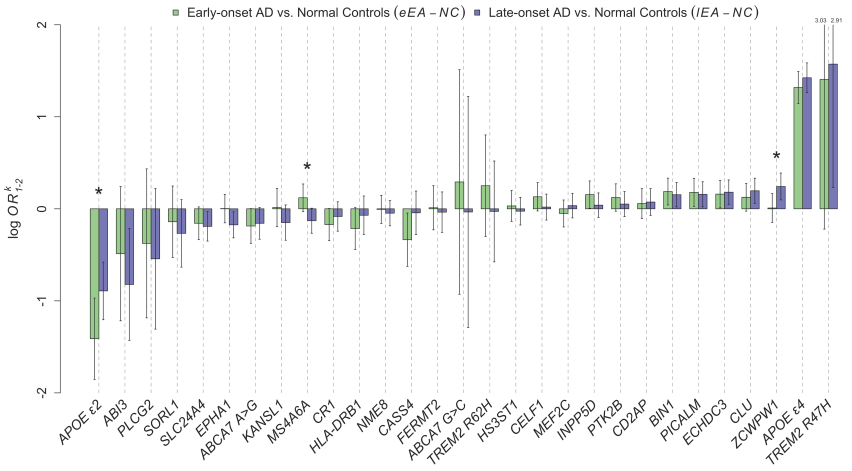


Figure 2.6: **Early onset AD vs normal controls and late onset AD vs normal controls.** Effect-sizes and 95% confidence intervals of a comparison of early onset AD cases ($\log OR^k_{eEA-NC}$, age at onset $\leq$ 65 years) and late-onset AD ($\log OR^k_{lEA-NC}$, age at onset > 65 years) with normal controls. For all the variants, the 95% confidence intervals overlapped. [*]: difference between $\log OR^k_{eEA-NC}$ and $\log OR^k_{lEA-NC}$ was significant (p<0.05, two-sample z-test).

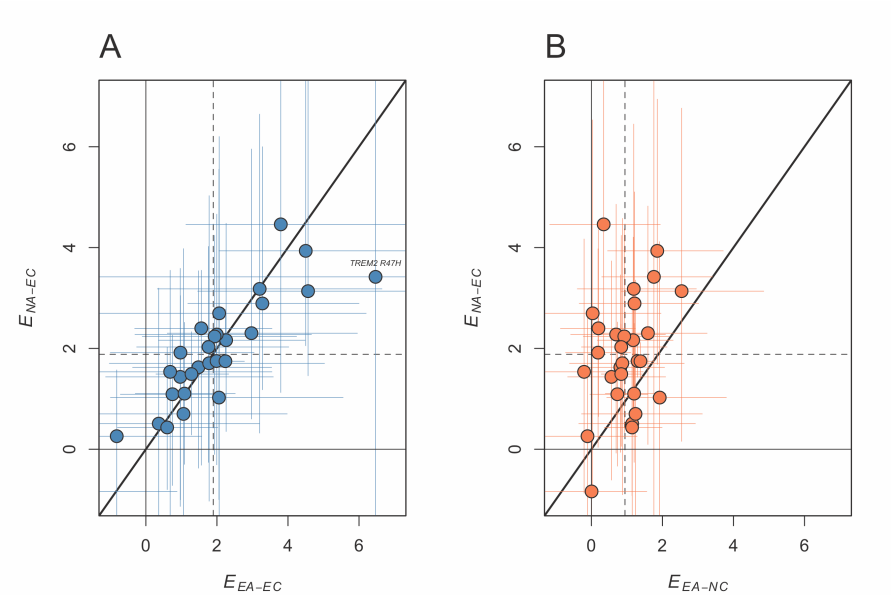


Figure 2.7: **Comparison of effect size changes at the variant level. A: Effect of using extreme AD cases vs. normal AD cases:** x-axis: E^k_{EA-EC}: Effect size changes from a comparison of the extreme cases and extreme (centenarian) controls relative to published effect sizes. Dashed line x-axis average effect-size increase E^k_{EA-EC} at 1.90 ± 0.29; y-axis: E^k_{NA-EC}: effect-size changes from a comparison normal AD cases with extreme (centenarian) controls relative to published effect sizes. Dashed line y-axis: average effect-size increase E^k_{NA-EC} at 1.88 ± 0.24. See Table 2.2 for E^k_{EA-NC} and Table S3 for E^k_{EA-NC} values. **B. Effect of using extreme cases vs. using extreme controls:** x-axis: effect-size changes of extreme AD cases vs. normal controls relative to published effect-sizes. Dashed line x-axis: average effect-size increase E^k_{EA-NC} at 0.94 ± 0.12. Y-axis: Variant effect-size change of normal AD cases vs. extreme controls relative to published effect-sizes. Dashed line y-axis: average effect-size increase E^k_{NA-EC} at 1.88 ± 0.24. See Table S3 for E^k_{EA-NC} and E^k_{NA-EC} values.

2.8 Supplementary Tables

Supplementary Tables can be accessed by scanning the following code or accessing the journal's website here.

References

[1] "2012 Alzheimer's disease facts and figures". In: *Alzheimer's & Dementia* 8.2 (Mar. 2012), pp. 131–168. ISSN: 15525260. DOI: 10.1016/j.jalz.2012.02.001.

[2] María M. Corrada et al. "Dementia incidence continues to increase with age in the oldest old: The 90+ study". In: *Annals of Neurology* 67.1 (Jan. 2010), pp. 114–121. ISSN: 03645134, 15318249. DOI: 10.1002/ana.21915.

[3] Margaret Gatz et al. "Role of genes and environments for explaining Alzheimer disease". In: *Archives of General Psychiatry* 63.2 (Feb. 2006), pp. 168–174. ISSN: 0003-990X. DOI: 10.1001/archpsyc.63.2.168.

[4] Jean-Charles Lambert et al. "Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease". In: *Nature Genetics* 41.10 (Oct. 2009), pp. 1094–1099. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.439.

[5] Denise Harold et al. "Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease". In: *Nature Genetics* 41.10 (Oct. 2009), pp. 1088–1093. ISSN: 1546-1718. DOI: 10.1038/ng.440.

[6] Sudha Seshadri et al. "Genome-wide analysis of genetic loci associated with Alzheimer disease". In: *JAMA* 303.18 (May 2010), pp. 1832–1840. ISSN: 1538-3598. DOI: 10.1001/jama.2010.574.

[7] Rahul S. Desikan et al. "Polygenic Overlap Between C-Reactive Protein, Plasma Lipids, and Alzheimer Disease". In: *Circulation* 131.23 (June 2015), pp. 2061–2069. ISSN: 1524-4539. DOI: 10.1161/CIRCULATIONAHA.115.015489.

[8] Rebecca Sims et al. "Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease". In: *Nature Genetics* 49.9 (Sept. 2017), pp. 1373–1384. ISSN: 1546-1718. DOI: 10.1038/ng.3916.

[9] Rita Guerreiro et al. "TREM2 variants in Alzheimer's disease". In: *The New England Journal of Medicine* 368.2 (Jan. 2013), pp. 117–127. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1211851.

[10] Thorlakur Jonsson et al. "Variant of TREM2 associated with the risk of Alzheimer's disease". In: *The New England Journal of Medicine* 368.2 (Jan. 2013), pp. 107–116. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1211103.

[11] Paul Hollingworth et al. "Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease". In: *Nature Genetics* 43.5 (May 2011), pp. 429–435. ISSN: 1546-1718. DOI: 10.1038/ng.803.

[12] Adam C. Naj et al. "Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease". In: *Nature Genetics* 43.5 (May 2011), pp. 436–441. ISSN: 1546-1718. DOI: 10.1038/ng.801.

[13] J. C. Lambert et al. "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease". In: *Nature Genetics* 45.12 (Dec. 2013), pp. 1452–1458. ISSN: 1546-1718. DOI: 10.1038/ng.2802.

[14] G. Jun et al. "A novel Alzheimer disease locus located near the gene encoding tau protein". In: *Molecular Psychiatry* 21.1 (Jan. 2016), pp. 108–117.

ISSN: 1476-5578. DOI: 10.1038/mp.2015.23.

[15] Stacy Steinberg et al. "Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease". In: *Nature Genetics* 47.5 (May 2015), pp. 445–447. ISSN: 1546-1718. DOI: 10.1038/ng.3246.

[16] W. J. Strittmatter et al. "Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease". In: *Proceedings of the National Academy of Sciences of the United States of America* 90.5 (Mar. 1993), pp. 1977–1981. ISSN: 0027-8424.

[17] E. H. Corder et al. "Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease". In: *Nature Genetics* 7.2 (June 1994), pp. 180–184. ISSN: 1061-4036. DOI: 10.1038/ng0694-180.

[18] Wiesje M. van der Flier and Philip Scheltens. "Amsterdam Dementia Cohort: Performing Research to Optimize Care". In: *Journal of Alzheimer's Disease* 62.3 (Mar. 2018). Ed. by George Perry, Jesus Avila, and Xiongwei Zhu, pp. 1091–1111. ISSN: 13872877, 18758908. DOI: 10.3233/JAD-170850.

[19] M. Huisman et al. "Cohort Profile: The Longitudinal Aging Study Amsterdam". In: *International Journal of Epidemiology* 40.4 (Aug. 2011), pp. 868–876. ISSN: 0300-5771, 1464-3685. DOI: 10.1093/ije/dyq219.

[20] Henne Holstege et al. "The 100-plus Study of Dutch cognitively healthy centenarians: rationale, design and cohort description". In: (Apr. 2018). DOI: 10.1101/295287.

[21] Wiesje M. van der Flier et al. "Optimizing patient care and research: the Amsterdam Dementia Cohort". In: *Journal of Alzheimer's disease: JAD* 41.1 (2014), pp. 313–327. ISSN: 1875-8908. DOI: 10.3233/JAD-132306.

[22] A. R. Varma et al. "Evaluation of the NINCDS-ADRDA criteria in the differentiation of Alzheimer's disease and frontotemporal dementia". In: *Journal of Neurology, Neurosurgery, and Psychiatry* 66.2 (Feb. 1999), pp. 184–188. ISSN: 0022-3050.

[23] D. Blacker et al. "Reliability and validity of NINCDS-ADRDA criteria for Alzheimer's disease. The National Institute of Mental Health Genetics Initiative". In: *Archives of Neurology* 51.12 (Dec. 1994), pp. 1198–1204. ISSN: 0003-9942.

[24] Anja Hviid Simonsen et al. "Recommendations for CSF AD biomarkers in the diagnostic evaluation of dementia". In: *Alzheimer's & Dementia* 13.3 (Mar. 2017), pp. 274–284. ISSN: 15525260. DOI: 10.1016/j.jalz.2016.09.008.

[25] Emiel O. Hoogendijk et al. "The Longitudinal Aging Study Amsterdam: cohort update 2016 and major findings". In: *European Journal of Epidemiology* 31.9 (Sept. 2016), pp. 927–945. ISSN: 0393-2990, 1573-7284. DOI: 10.1007/s10654-016-0192-0.

[26] Sayantan Das et al. "Next-generation genotype imputation service and methods". In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1284–1287. ISSN: 1546-1718. DOI: 10.1038/ng.3656.

[27] Shane McCarthy et al. "A reference panel of 64,976 haplotypes for genotype imputation". In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1279–1283. ISSN: 1546-1718. DOI: 10.1038/ng.3643.

[28] Jared O'Connell et al. "A general approach for haplotype phasing across the full spectrum of relatedness". In: *PLoS genetics* 10.4 (Apr. 2014), e1004234. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004234.

[29] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (Oct. 2015), pp. 68–74. ISSN: 1476-4687. DOI: 10.1038/nature15393.

[30] Carl A. Anderson et al. "Data quality control in genetic case-control association studies". In: *Nature Protocols* 5.9 (Sept. 2010), pp. 1564–1573. ISSN: 1750-2799. DOI: 10.1038/nprot.2010.116.

[31] Alkes L. Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature Genetics* 38.8 (Aug. 2006), pp. 904–909. ISSN: 1061-4036. DOI: 10.1038/ng1847.

[32] Alkes L. Price et al. "New approaches to population stratification in genome-wide association studies". In: *Nature Reviews. Genetics* 11.7 (July 2010), pp. 459–463. ISSN: 1471-0064. DOI: 10.1038/nrg2813.

[33] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *American Journal of Human Genetics* 81.3 (Sept. 2007), pp. 559–575. ISSN: 0002-9297. DOI: 10.1086/519795.

[34] R. Core Team. "R: A language and environment for statistical computing." In: R Foundation for Statistical Computing, Vienna, Austria. ().

[35] Caroline Van Cauwenberghe, Christine Van Broeckhoven, and Kristel Sleegers. "The genetic landscape of Alzheimer disease: clinical implications and perspectives". In: *Genetics in Medicine* 18.5 (May 2016), pp. 421–430. ISSN: 1098-3600, 1530-0366. DOI: 10.1038/gim.2015.117.

[36] A. M. Saunders et al. "Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease". In: *Neurology* 43.8 (Aug. 1993), pp. 1467–1472. ISSN: 0028-3878.

[37] Aldi T. Kraja et al. "Genetic analysis of 16 NMR-lipoprotein fractions in humans, the GOLDN study". In: *Lipids* 48.2 (Feb. 2013), pp. 155–165. ISSN: 1558-9307. DOI: 10.1007/s11745-012-3740-8.

[38] Angela R. Brooks-Wilson. "Genetics of healthy aging and longevity". In: *Human Genetics* 132.12 (Dec. 2013), pp. 1323–1338. ISSN: 1432-1203. DOI: 10.1007/s00439-013-1342-z.

[39] Jacob vB Hjelmborg et al. "Genetic influence on human lifespan and longevity". In: *Human Genetics* 119.3 (Apr. 2006), pp. 312–321. ISSN: 0340-6717. DOI: 10.1007/s00439-006-0144-y.

[40] Peter K. Joshi et al. "Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity". In: *Nature Communications* 8.1 (Dec. 2017). ISSN: 2041-1723. DOI: 10.1038/s41467-017-00934-5.

[41] Seungjin Ryu et al. "Genetic landscape of APOE in human longevity revealed by high-throughput sequencing". In: *Mechanisms of Ageing and Development* 155 (Apr. 2016), pp. 7–9. ISSN: 00476374. DOI: 10.1016/j.mad.2016.02.010.

[42] Linda Broer et al. "GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy". In: *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 70.1 (Jan. 2015), pp. 110–118. ISSN:

1758-535X. DOI: 10.1093/gerona/glu166.

[43] Paola Sebastiani et al. "Four Genome-Wide Association Studies Identify New Extreme Longevity Variants". In: *The Journals of Gerontology: Series A* 72.11 (Oct. 2017), pp. 1453–1464. ISSN: 1079-5006, 1758-535X. DOI: 10.1093/gerona/glx027.

[44] Jenny Lord, Alexander J. Lu, and Carlos Cruchaga. "Identification of rare variants in Alzheimerâ€™s disease". In: *Frontiers in Genetics* 5 (Oct. 2014). ISSN: 1664-8021. DOI: 10.3389/fgene.2014.00369.

[45] Riccardo E. Marioni et al. "GWAS on family history of Alzheimer's disease". In: *Translational Psychiatry* 8.1 (Dec. 2018), p. 99. ISSN: 2158-3188. DOI: 10.1038/s41398-018-0150-6.

[46] Iris E. Jansen et al. "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk". In: *Nature Genetics* 51.3 (Mar. 2019), pp. 404–413. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-018-0311-9.

3. Resilience against dementia

Immune response and endocytosis pathways are associated with the resilience against Alzheimer's Disease

Niccolo' Tesi, Sven J. van der Lee, Marc Hulsman, Iris E. Jansen, Najada Stringa, Natasja M. van Schoor, Martijn Huisman, Philip Scheltens, Wiesje M. van der Flier, Marcel J.T. Reinders, and Henne Holstege

This chapter was published in *Translational Psychiatry*
https://doi.org/10.1038/s41398-020-01018-7

Abstract

Developing Alzheimer's disease (AD) is influenced by multiple genetic variants that are involved in five major AD-pathways. Per individual, these pathways may differentially contribute to the modification of the AD-risk. The pathways involved in the *resilience* against AD have thus far been poorly addressed. Here, we investigated to what extent each molecular mechanism associates with (*i*) the increased risk of AD and (*ii*) the resilience against AD until extreme old age, by comparing pathway-specific polygenic risk scores (pathway-PRS). We used 29 genetic variants associated with AD to develop pathway-PRS for five major pathways involved in AD. We developed an integrative framework that allows multiple genes to associate with a variant, and multiple pathways to associate with a gene. We studied pathway-PRS in the Amsterdam Dementia Cohort of well-phenotyped AD patients (*N*=1,895), Dutch population controls from the Longitudinal Aging Study Amsterdam (*N*=1,654) and our unique 100-plus Study cohort of cognitively healthy centenarians who avoided AD (*N*=293). Last, we estimated the contribution of each pathway to the genetic risk of AD in the general population. All pathway-PRS significantly associated with increased AD-risk and (in the opposite direction) with resilience against AD (except for angiogenesis, $p<0.05$). The pathway that contributed most to the overall modulation of AD-risk was β-amyloid metabolism (29.6%), which was driven mainly by *APOE*-variants. After excluding *APOE* variants, all pathway-PRS associated with increased AD-risk (except for angiogenesis, $p<0.05$), while specifically immune response ($p=0.003$) and endocytosis ($p=0.0003$) associated with resilience against AD. Indeed, the variants in these latter two pathways became the main contributors to the overall modulation of genetic risk of AD (45.5% and 19.2%, respectively). The genetic variants associated with the resilience against AD indicate which pathways are involved with maintained cognitive functioning until extreme ages. Our work suggests that a favorable immune response and a maintained endocytosis pathway might be involved in general neuro-protection, which highlight the need to investigate these pathways, next to β-amyloid metabolism.

3.1 Introduction

Owing changes in lifestyle and advances in healthcare, life expectancy has greatly increased during the last century.[1] A consequence of an increased fraction of aged individuals in the population is the increased prevalence of age-related diseases. A major contribution to poor health and disability at old age is cognitive decline due to Alzheimer's disease (AD).[2] The incidence of AD increases exponentially with age and reaches ~40% per year at 100 years, making it one of the most prevalent diseases in the elderly.[3] Yet, a small proportion of the population (<0.1%) avoids the disease, reaching at least 100 years while maintaining a high level of cognitive health.[4] Both the development and the resilience against AD are determined by a combination of beneficial and harmful environmental and genetic factors that is unique for each individual.[1, 5, 6] Thus far, large collaborative genome-wide association studies (GWAS) have discovered common genetic variants associated with a small modification of the risk of AD.[7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20] Of these, the alleles that encompass the *APOE* gene explain the largest proportion of the risk to develop or the chance to escape AD. We previously showed that those who avoided cognitive decline until extreme ages (cognitively healthy centenarians) were relatively depleted with genetic variants associated with an increased risk of AD.[21] However, the degree of depletion of these variants in the genomes of cognitively healthy centenarians relative to the middle-aged healthy individuals was not constant, which might point towards a differential impact of associated biological pathways on either avoiding or developing AD. This led us to hypothesize that an individuals' chance to develop AD or to being resilient against AD may be determined by pathway-specific risk. Previous studies indicated that five specific biological pathways associate strongly with AD risk: immune response, β-amyloid metabolism, cholesterol/lipid dysfunction, endocytosis and angiogenesis.[22, 23, 24, 25, 26, 27] However, the extent to which different pathways contribute to the polygenic risk of AD is unknown. The degree to which a pathway contributes to the individual risk can be studied with pathway-specific polygenic risk scores (PRS).[28, 29] In a typical polygenic risk score, the effect-sizes of all genetic variants that significantly associate with a trait are combined.[30] In a pathway-specific PRS, additional information is necessary: (*i*) the association of genetic variants to genes, and (*ii*) the association of genes to pathways. Previous studies of pathway-PRS in AD approached these challenges using the closest gene for variant mapping. For this, a 1:1 relationship between variants and genes is assumed,

however, as AD-associated variants are mostly intronic or intergenic, the closest gene is not necessarily the gene affected by the variant. Additionally, different databases often have different functional annotations of genes, and this uncertainty was previously not taken into account when constructing pathway-PRS.[28, 29]

An accurate mapping of the genetic risk of AD conferred by specific molecular pathways may lead to a greater comprehension of individual AD subtypes and might represent a first important step for the development of targeted intervention strategies and personalized medicine.[31] Here, we propose a novel integrative framework to construct pathway-PRS for the five major pathways suggested to be involved in AD. We then tested whether specific pathways differentially contributed to the risk of AD as well as to the chance of avoiding AD until extreme old ages. Finally, we estimated the contribution of each pathway to the polygenic risk of AD in the general (healthy middle-aged) population.

3.2 Methods

3.2.1 Populations

Population subjects are denoted by *P*: they consist of a representative Dutch sample of 1,779 individuals aged 55-85 years from the Longitudinal Aging Study Amsterdam (LASA).[32, 33] Patients diagnosed with AD are denoted by *A*. The patients are either clinically diagnosed probable AD patients from the Amsterdam Dementia Cohort (N=1,630) or pathologically confirmed AD patients from the Netherlands Brain Bank (N=436).[34, 35, 36] Escapers of AD are denoted by *C*: these are 302 cognitively healthy centenarians from the 100-plus Study cohort. This study includes individuals who can provide official evidence for being aged 100 years or older and self-report to be cognitively healthy, which is confirmed by a proxy.[4] All participants and/or their legal representatives provided written informed consent for participation in clinical and genetic studies. The Medical Ethics Committee of the Amsterdam UMC (METC) approved all studies.

3.2.2 Genotyping and imputation

We selected 29 common genetic variants (minor allele frequency >1%) for which a genome-wide significant association with clinically identified AD cases was found (Table S1).[7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 37, 38, 39] We genotyped all individuals using Illumina Global Screening Array (GSAsharedCUSTOM_20018389_A2) and applied established quality control measures.[40] Briefly, we used high-quality genotyping in all individuals (individual call rate >98%, variant call rate >98%) and Hardy-Weinberg equilibrium-departure was considered significant at p <1x10^{-6}. Genotypes were prepared for imputation using provided scripts (HRC-1000G-check-bim.pl).[41] This script compares variant ID, strand and allele frequencies to the haplotype reference panel (HRC v1.1, April 2016).[41] Finally, all autosomal variants were submitted to the Michigan imputation server (`https://imputationserver.sph.umich.edu`). The server uses SHAPEIT2 (v2.r790) to phase data and imputation to the reference panel (v1.1) was performed with Minimac3. Variant-genotypes of total of 1,779 population subjects, 302 centenarians and 2,052 AD cases passed quality control. Prior to analysis, we excluded individuals of non-European ancestry (N_C = 2, N_P = 63 and N_A = 94 based on 1000Genomes clustering)[42] and individuals with a family relation (N_C = 7, N_P = 62 and N_A = 63, identity-by-descent >0.3), leaving 1,654 population subjects, 293 cognitively healthy centenarians and 1,895 AD cases for the analyses.

3.2.3 Polygenic risk score

To calculate the personal polygenic risk scores, or the genetic risk of AD that affects a single individual, the effect-sizes of all genetic variants that significantly associate with AD are combined. Formally, a PRS is defined as the sum of trait-associated alleles carried by an individual across a defined set of genetic loci, weighted by effect-sizes estimated from a GWAS.[30] We constructed a polygenic risk score (PRS) using 29 variants that were previously associated with AD. As weights for the PRS, we used the variant effect-sizes (log of odds ratio) as published in large GWAS of AD (Table S1). Given a subject s, the PRS is defined as:

$$PRS^s = \sum_k^K (dos_k^s * \beta_k) \tag{3.1}$$

where K is the full set of variants, dos_k^s is the allele dosage from the (imputed) genotype of variant k in subject s and β_k is the effect size as determined in the largest published AD case-control GWAS (Table S1).

3.2.4 Mapping variants to pathways

We studied the five pathways implicated in AD: immune response, β-amyloid metabolism, cholesterol/lipid dysfunction, endocytosis and angiogenesis.[22, 23, 24, 25, 43, 44] For these pathways we developed the variant-pathway mapping M_p^k, which represents the degree of involvement of a given variant in the pre-selected pathways. To generate this value, we (*i*) associated genetic variants to genes (variant-gene mapping), (*ii*) associated genes to pathways (gene-pathway mapping) and (*iii*) combined these mappings in the variant-pathway mapping.

Variant-gene mapping

The association of a variant with a specific gene is not straight-forward as the closest gene is not necessarily the gene affected by the variant. The two most recent and largest GWAS of AD addressed the relationship between genetic variants and associated genes applying two independent methods.[20, 19] Briefly, one study used (*i*) gene-based annotation, (*ii*) expression-quantitative trait loci (eQTL) analyses, (*iii*) gene cluster/pathway analyses, and (*iv*) differential gene expression analysis between AD cases and healthy controls.[19] The other study integrated (*i*) positional mapping, (*ii*) eQTL gene-mapping, and (*iii*) chromatin interaction as implemented in the tool Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA).[20, 45]

The list of genes most likely affected by each variant was obtained from both studies and used to derive a weighted mapping for each genetic variant k to one or more genes g, m_g^k, denoted as the variant-gene mapping weight. This weight was calculated by counting the number of times a variant k was associated with gene g across the two studies and dividing this by the total number of genes associated with the variant (Table S2). For variants in/near *CR1, PILRA, PLCG2, ABCA7* and *APOE*, we assumed the culprit gene as known, and we assigned a 1:1 relationship between the variant and the gene (Table S2).

Gene-pathway mapping

Each gene from the variant-gene mapping was classified into the pre-defined set of pathways integrating four sources of information:

- Gene-sets from the unsupervised pathway enrichment analysis within MAGMA statistical framework from *Kunkle et al.*,[19] in which the authors identified 9 significant pathways (coupled with the genes involved in each pathway), which we mapped to 3 of the 5 pathways of interest (Table S3);
- Associated genes from Gene-ontology (GO, from AmiGO 2 version 2.5.12, released on 2018-04) terms resembling the 5 pathways of interest within the biological processes tree (including all child-terms) (Table S4);[46, 47]
- Gene-sets derived from an unsupervised functional clustering analysis within DAVID (v6.8, released on 2016-10):[48, 49] the gene-set from the variant-gene mapping was used to obtain 12 functional clusters which were then mapped to the 5 pre-selected pathways using a set of keywords (Table S5 and Table S6);
- Gene-pathway associations from a recent review concerning the genetic landscape of AD (Table S7);

By counting the number of times each gene was associated to each pathway according to these sources, and dividing by the total number of associations per gene, we obtained a weighted mapping of each gene g to one or more pathways p, w_p^g, denoted as the gene-pathway mapping weight (Table S8 and Table S9). In case the gene-pathway mapping could not be calculated (*i.e.* there was no mapping to any of the pathways of consideration), we excluded the gene from further analyses (Table S8 and Table S9).

Variant-pathway mapping

To associate variants with pathways, we combined the *variant-gene mapping*

and the *gene-pathway mapping*. Given a variant k, mapping to a set of genes G, and a pathway p, we define the weight of the variant to the pathway (M_p^k) as:

$$M_p^k = \sum_g^G (m_g^k * w_p^g) \tag{3.2}$$

where m_g^k is the variant-gene mapping weight of variant k to gene g, and w_p^g is the gene-pathway mapping weight of gene g to pathway p. In this way, for each variant, we calculated a score indicative of the involvement of the variant in each of the five pathways (*variant-pathway mapping*, Table S10). For some variants no *variant-pathway mapping* was possible. We marked these variants as unmapped (Table S10).

3.2.5 Pathway-specific polygenic risk score

For the pathway-specific polygenic risk score (pPRS), we extended the definition of the PRS by adding as multiplicative factor the *variant-pathway mapping* weight of each variant. Given a sample s and a pathway p, we defined the pPRS as:

$$pPRS_p^s = \sum_k^K (dos_k^s * \beta_k * M_k^p) \tag{3.3}$$

where M_k^p is the *variant-pathway mapping* of variant k to pathway p.

3.2.6 Association of PRSs in the three cohorts

We calculated the polygenic risk score (PRS) and pathway-PRS (pPRS) for the population subjects, the AD cases and the cognitively healthy centenarians (*P*, *A* and *C*, respectively). Prior to analyses, the PRSs of all three populations were combined together and were scaled (μ=0, σ=1). We then investigated the influence of *APOE*, gender and age on the risk scores: we calculated the PRSs and pPRSs with and without the two *APOE* variants and we correlated the resulting (p)PRSs with sex, age (age at inclusion for controls, age at onset for cases) and population substructure components. To inspect the differential contributions of the risk scores to AD development or resilience against AD, we calculated (*i*) the association of the risk scores (PRS and pPRS) with AD status by comparing AD cases and population subjects (*A vs. P*), and (*ii*) the association of the risk scores with resilience against AD by comparing cognitively healthy centenarians and population subjects (*C*

vs. P comparison). For the associations, we used logistic regression models with the PRS and pPRS as predictors, adjusting for population substructure (principal components 1-5). Resulting effect-sizes (log of odds ratio) can be interpreted as the odds ratio difference per one standard deviation (SD) increase in the PRS, with a corresponding estimated 95% confidence intervals (95% CI). Association analyses of the (p)PRS in the three population were also stratified by sex. Last, we verified the classification performances of the single variants as well as the (p)PRS by calculating the area under the *ROC* curve for classification of AD and resilience against AD.

3.2.7 Resilience against AD vs. increased AD-risk

To further investigate the relationship between the effect of each pathway on AD and on resilience against AD, we calculated the change in effect-size. This corresponds to the ratio between the effect-size of the association with resilience against AD (log of odds ratios of *C vs. P* comparison) and the effect-size of the association with AD (log of odds ratios of *A vs. P* comparison). We calculated the change in effect-size for the pPRS including and excluding *APOE* variants. We estimated 95% confidence intervals for the effect-size ratios by sampling, and we tested for significant difference between the change in effect-size including and excluding *APOE* variants (respectively for each of pPRS) using *t*-test. A value for the change in effect-size of 1 indicates a similar effect on increased risk of AD and resilience against AD. Although a value for the change in effect size is unknown a priori, since all variants considered are selected to be associated with AD, a value <1 is expected (*i.e.* a larger effect on AD than on resilience against AD).

3.2.8 Contribution of each pathway to polygenic risk of AD

We estimated the contribution of each pathway to the genetic risk of AD in the general population: this equals to the variance explained by each of the pre-selected pathways to the genetic risk of AD. Mathematically, this is the ratio between the variance of each pathway-PRS and the variance of the combined PRS as calculated in the individuals of the general population. As such, it is a function of the variant-pathway mapping, the effect-size (log of odds ratio) of the variants, and the variant frequencies. Given a variant k and the relative *variant-pathway mapping* M_p^k, we defined the percentage P of the risk explained by each pathway p as:

$$P^p = \frac{\sum_k^K (M_k^p * \beta_k^2 * MAF_k * (1 - MAF_k)}{\sum_k^K \beta_k^2 * MAF_k * (1 - MAF_k))} \tag{3.4}$$

where β_k is the variant effect-size from literature, and $MAF_k * (1 - MAF_k)$ is the variance of a Bernoulli random variable that occurs with probability MAF_k, *i.e.* the minor allele frequency of each variant k in our cohort of population subjects. Here, M_p^k is interpreted as the probability that variant k belongs to pathway p. Importantly, for each variant, $\sum_p^P M_p^k = 1$, so that each variant contributes equally, yet differentially at the level of each pathway. This means that the variance of a variant is only counted once, even if the variant contributes to multiple pathways. When calculating the contributions of each pathway, we also considered variants with missing variant-pathway mapping. For these variants, the variant-pathway mapping was set to 1 for an unmapped pathway. Together, the pathway PRS variances sum to the total PRS variance.

3.2.9 Implementation

We performed quality control of genotype data as well as population stratification analysis and relatedness analysis with PLINK (v2.0). All subsequent analyses were performed with R (v3.5.2), Bash and Python (v2.7.14) scripts. We provide a R script to construct pPRS and PRS using our variant-pathway annotation and user's genotypes. In addition, all the scripts we used to perform the analyses can be found at `https://github.com/TesiNicco/pathway-PRS`.

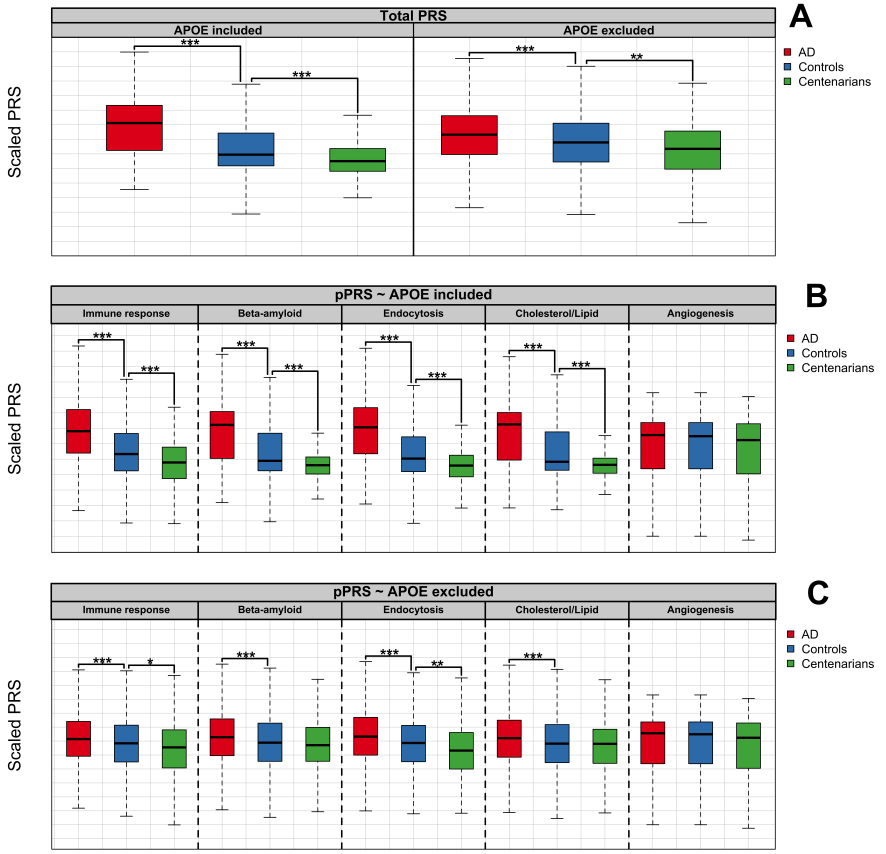


Figure 3.1: **Boxplots of PRS and pPRS in the different settings. A.** The PRS including all the 29 known AD-associated variants, with and without *APOE* variants. As weight for the PRS, we used published variant effect-sizes (Table S1). **B.** and **C.** The pPRS for each of the selected molecular pathways, including and excluding *APOE* variants, respectively. For all plots, risk scores were calculated for AD cases, population subjects and cognitively healthy centenarians. Then, risk scores were compared between (*i*) AD cases and population subjects (*A vs. P* comparison) and (*ii*) cognitively healthy centenarians and population subjects (*C vs. P* comparison). For representation, we scaled all PRS and pathway-PRS to be μ=0 and σ=1. For the comparison, we used logistic regression models with risk scores as predictors. Annotation: ***, *p*-value of association <5x10^{-6}; **, *p*-value of association <0.0005; *, *p*-value of association <0.05.

3.3 Results

After quality control of the genetic data, we included 1,654 population subjects (with mean age at inclusion 62.7 ± 6.4, 53.2% females), 1,895 AD cases (with mean age at onset 69.2 ± 9.9, 56.4% females), and 293 cognitively healthy centenarians (with mean age at inclusion 101.4 ± 1.3, 72.6% females) (*P*, *A* and *C* respectively).

3.3.1 Polygenic risk scores associate with AD and escape from AD

To each subject, we assigned a PRS representative of all 29 AD-associated variants, including and excluding *APOE* variants. We found that the PRS, when including *APOE* variants, significantly associated with an increased risk of AD and, in the opposite direction, with increased chance of resilience against AD (*A vs. P*: OR=2.61, 95% CI=[2.40-2.83], p=8.4x10^{-113} and *C vs. P*: OR=0.54, 95% CI=[0.45-0.65], p=1.1x10^{-10} (Figure 3.1 and Table S11). When excluding *APOE* variants, the PRS was still significantly associated with an increased risk of AD and, in the opposite direction, with increased risk of resilience against AD (*A vs. P*: OR=1.30, 95% CI=[1.22-1.40], p=3.1x10^{-14} and *C vs. P*: OR=0.78, 95% CI=[0.69-0.89], p=2.4x10^{-4}) (Figure 3.1, and Table S11).

3.3.2 Pathway-specific PRS associate with AD and escape from AD

We annotated the 29 AD-associated genetic variants to 5 selected pathways (Figure 3.2). According to our variant-gene mapping, the 29 AD-associated variants mapped to 110 genes (Table S8). The number of genes associated with each variant ranged from 1 (*e.g.* for variants in/near *CR1, PILRA, SORL1, ABCA7, APOE* or *PLCG2*, to 30 (a variant in the gene-dense region within the HLA region) (Figure 3.2 and Table S8). We were able to calculate the gene-pathway mapping weight for 69 genes (Table S9). The remaining 41 genes were not mapped to the 5 pathways. In total, we calculated the variant-pathway mapping for 23 loci to at least one of the pre-selected biological pathways (Figure 3.2 and Table S10).

We then calculated the pPRS for each pathway in population subjects, AD cases and cognitively healthy centenarians including and excluding *APOE* variants (Figure 3.1B and Figure 3.1C). The number of variants that contributed to each pPRS was 19 for immune response, 11 for β-amyloid metabolism, 19 for endocytosis, 8 for cholesterol/lipid dysfunction and 4 for angiogenesis pathways (Table S10 and Table S11). Overall, the pPRS (including and excluding the *APOE* variants) positively and significantly correlated with each other and with the overall PRS (Figure 3.5), and did not

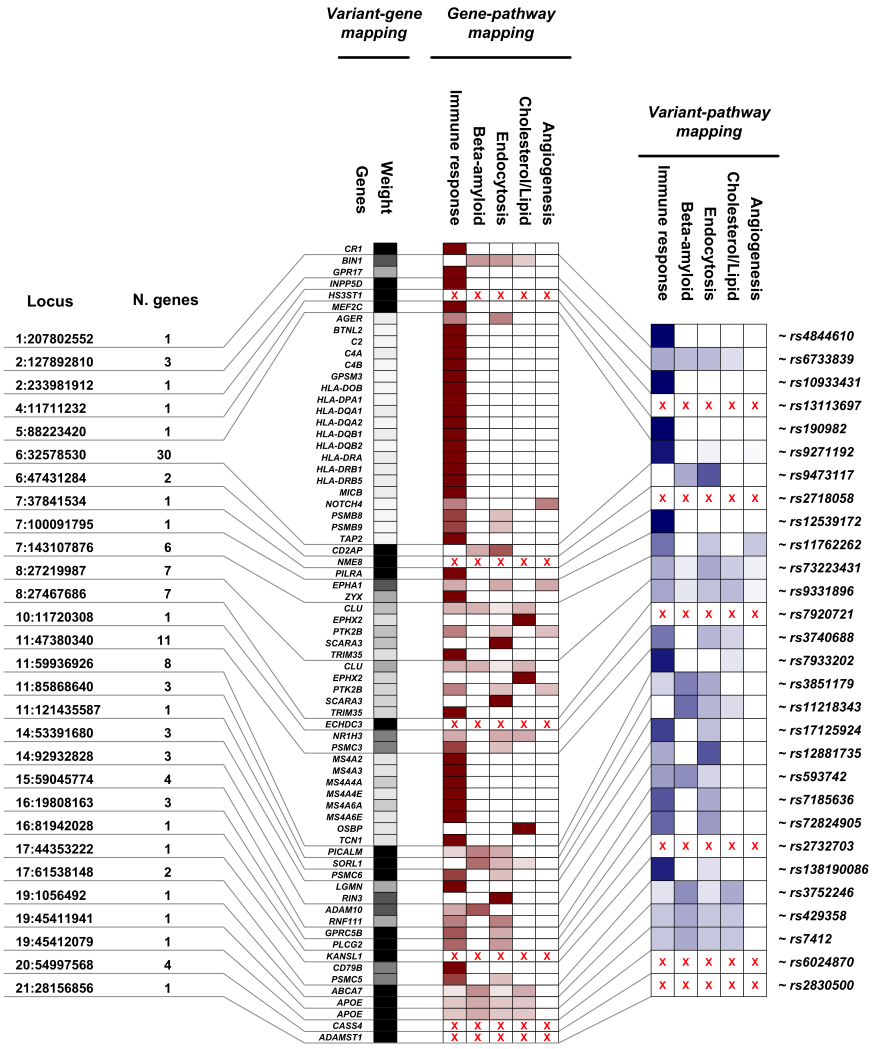


Figure 3.2: **Variant-pathways mapping.** *Locus*: chromosome and position of the AD-associated genetic variants (coordinates are with respect to GRCh37). *N.genes*: total number of genes associated with each variant according to variant-gene mapping. *Variant-gene mapping*: *Genes*: all genes with at least one annotation to the 5 selected molecular pathways associated with AD. *Weight*: the weight of the variant-gene mapping. *Gene-pathway mapping: Immune response, Beta-amyloid, Endocytosis, Cholesterol/lipid, Angiogenesis*: the weight of each molecular pathway at the gene level. *Variant-pathway mapping*: summarization of each variant's effect after combining variant-gene and gene-pathway mappings. Red crosses indicate unmapped genes.

correlate with gender and age (Figure 3.5).

When including *APOE* variants, the pPRSs of all pathways (except for angiogenesis) significantly associated with increased risk of AD, independently from gender (*A vs. P*, immune response: OR=2.15, 95% CI=[1.99-2.32], p=2.0x10^{-80}; β-amyloid metabolism: OR=2.52, 95% CI=[2.32-2.73], p=7.8x10^{-109}; endocytosis: OR=2.55, 95% CI=[2.35-2.77], p=1.7x10^{-109}; cholesterol/lipid dysfunction: OR=2.55, 95% CI=[2.35-2.76], p=2.1x10^{-110}; angiogenesis: OR=1.05, 95% CI=[0.98-1.12], p=0.134) (Figure 3.1B, Table S11, Figure 3.6 and Table S12). The association of pPRSs with increased chance of being resilient against AD was in the opposite direction for all pathways, and the association was significant for all pathways except for angiogenesis (*C vs. P*, immune response: OR=0.64, 95% CI=[0.54-0.74], p=1.4x10^{-8}; β-amyloid metabolism: OR=0.59, 95% CI=[0.49-0.71], p=2.7x10^{-8}; endocytosis: OR=0.55, 95% CI=[0.46-0.66], p=1.3x10^{-10}; cholesterol/lipid dysfunction: OR=0.58, 95% CI=[0.48-0.70], p=1.8x10^{-8}; angiogenesis: OR=0.90, 95% CI=[0.79-1.01], p=0.078) (Figure 3.1B, Table S11). Directions of effects were consistent in both males and females, but the significance of associations was reduced due to stratification (Table S12 and Figure 3.5). When excluding *APOE* variants, the pPRSs of all pathways (except for the angiogenesis) was still significantly associated with increased risk of AD without specific gender effects (*A vs. P*, immune response: OR=1.19, 95% CI=[1.11-1.27], p=5.5x10^{-7}; β-amyloid metabolism: OR=1.19, 95% CI=[1.12-1.28], p=2.0x10^{-7}; endocytosis: OR=1.27, 95% CI=[1.19-1.36], p=2.8x10^{-12}; cholesterol/lipid dysfunction: OR=1.18, 95% CI=[1.11-1.27], p=7.5x10^{-7}; angiogenesis: OR=1.05, 95% CI=[0.98-1.12], p=0.134) (Figure 3.1C, Table S11, Figure 3.6 and Table S12). The association of pPRSs with increased chance of being resilient against AD was in the opposite direction for all pathways, yet the association was significant only for the immune response and the endocytosis pPRS (*C vs. P*, immune response: OR=0.82, 95% CI=[0.72-0.94], p=0.003; β-amyloid metabolism: OR=0.91, 95% CI=[0.80-1.03], p=0.131; endocytosis: OR=0.79, 95% CI=[0.70-0.90], p=0.0003; cholesterol/lipid dysfunction: OR=0.91, 95% CI=[0.80-1.03], p=0.145; angiogenesis: OR=0.90, 95% CI=[0.79-1.01], p=0.078) (Figure 3.1C and Table S11). In the sex-stratified analysis, females reported consistent direction of effects and significant associations of immune response and endocytosis pathways, while in males the direction was consistent for immune response, endocytosis and angiogenesis pathways, and it was opposite for β-amyloid metabolism and cholesterol/lipid dysfunction (yet not significant) (Figure 3.6 and Table S12). We note that apart from *APOE* variants (for which we stratified the analyses for), there was no major driver in the pPRS as well as the single-variant

associations (Figure 3.7 and Figure 3.8).

3.3.3 Comparison of effect on AD and escaping AD

To further evaluate the association of the pPRSs with AD and with resilience against AD, we compared, for each pPRS, the reciprocal effect size associated with resilience against AD with the effect size associated with increased risk of AD (change in effect size, Figure 3.3A). When including *APOE* variants, the change in effect-size was <1 for all pathways (except for the angiogenesis pathway) (Figure 3.3B). This is expected as the effect-size of *APOE* variants on causing AD is much larger than its effect on resilience against AD (Figure 3.3A). When excluding *APOE* variants, the change in effect-size was still <1 for β-amyloid metabolism and cholesterol/lipid metabolism (respectively 0.54 and 0.58), but it approximated 1 for endocytosis (0.96) and it was larger than 1 for the immune response and angiogenesis (respectively 1.12 and 2.15) (Figure 3.3B). Interestingly, we found that the relative effect-size for immune response and endocytosis excluding *APOE* variants was significantly higher than that including *APOE* variants ($p<2.1\text{x}10^{-197}$ and $p<8.9\text{x}10^{-180}$ respectively), suggesting a larger effect on resilience against AD compared to AD-risk for these pathways, specifically when excluding *APOE* variants (Figure 3.3B).

3.3.4 Contributions of each pathway to the polygenic risk of AD

Finally, we estimated the relative contribution of each pathway to the polygenic risk of AD in the general population. This is indicative of the degree of involvement of each pathway to the total polygenic risk of AD, and as such it is based on out variant-pathway mapping. Including *APOE* variants, the contribution of the pathways to the total polygenic risk of AD was 29.6% for β-amyloid metabolism, 26.6% for immune response, 21.6% for endocytosis, 19.5% for cholesterol/lipid dysfunction, 0.3% for angiogenesis and 2.3% for the unmapped variants (Figure 3.4A). When we excluded *APOE* variants, the contribution of the pathways to the total polygenic risk of AD was 45.5% for immune response, 19.2% for endocytosis, 13.7% for β-amyloid metabolism, 8% for cholesterol/lipid dysfunction, 1.4% for angiogenesis and 12.3% for the unmapped variants (Figure 3.4B).

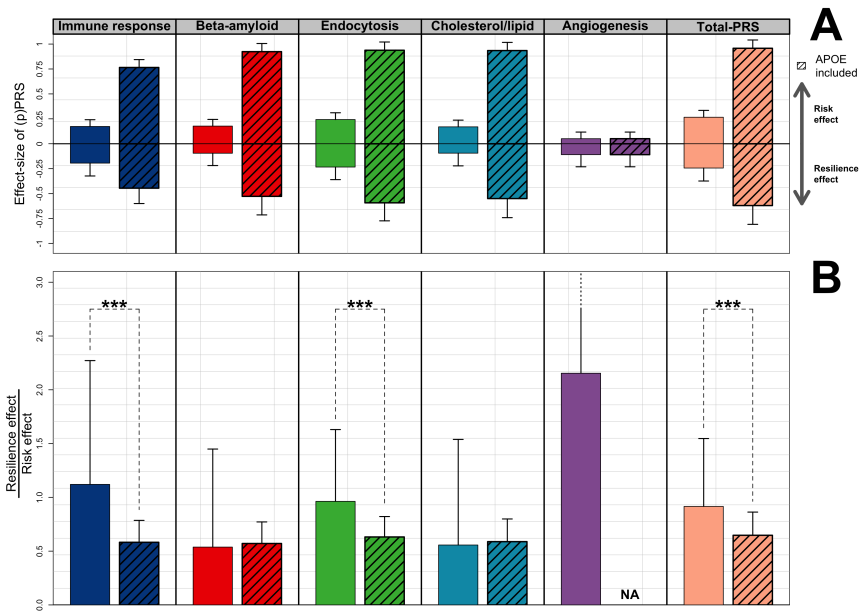


Figure 3.3: **Change in effect-size between association with escaping AD and causing AD for the five pPRSs. A** The effect-sizes (log of odds ratio) and the relative 95% confidence intervals of the association of the (p)PRS with both AD-risk and resilience against AD, grouped by pathway. **B.** Each bar represents the ratio between the effect-size of the association with escaping AD (Resilience effect in **A**) and with causing AD (Risk effect in **A**), respectively with and without *APOE* variants. Ratios larger than 1 are then indicative of larger effect-size on resilience against AD compared to AD-risk. We then compared the change in effect-size for each pathway when including and excluding *APOE* variants using *t*-tests. Annotation: ***, *p*-value of association <5x10^{-6}; **, *p*-value of association < 0.0005; *, *p*-value of association < 0.05.

3.4 Discussion

In this work, we studied 29 common genetic variants known to associate with AD using polygenic risk scores and pathway-specific polygenic risk scores. As expected, we found that a higher PRS for AD was associated with a higher risk of AD. Previous studies showed that polygenic risk score of AD not only associated with increased risk of AD, but also with neuropathological hallmarks of AD, lifetime risk and the age at onset in both *APOE* ϵ4 carriers and non-carriers.[28, 29, 50, 51, 52, 53, 54] We now add that, using our unique cohort of cognitively healthy centenarians, the PRS for AD also

associates with resilience against AD at extremely old ages. This adds further importance to the potentiality of using PRS and *APOE* genotype in a clinical setting.[51, 50, 53, 55] In addition, our analyses suggest that the long-term preservation of cognitive health is associated with the selective survival of individuals with the lowest burden of risk-increasing variants or, vice versa, the highest burden of protective variants. Using an innovative approach, we studied five pathways previously found to be involved in AD as well as the contribution of these pathways to the polygenic risk of AD. We showed that all pathways-PRS except angiogenesis associate with increased AD risk, both including and excluding *APOE* variants and independently from gender. When we studied the association of pathways-PRS with resilience against AD until extreme old ages, we found that, as expected, the enrichment of the protective *APOE* $\epsilon 2$ allele and the depletion of the risk-increasing *APOE* $\epsilon 4$ allele represented a major factor in avoiding AD. However, when excluding the two *APOE* variants, only immune response and endocytosis significantly associated with an increased chance to be resilient against AD. Interestingly, both pathways had a larger or similar effect on resilience against AD-resilience compared to developing AD, suggesting that these pathways might be involved in general neuro-protective functions. Based on the variant effect size, variant frequency and our variant-pathway mapping, we found that the β-amyloid metabolism (29.6%) followed by immune response (26.6%) were the major contributors to general modification of AD-risk. After excluding *APOE* variants, according to our analysis, immune response (45.5%) and endocytosis (19.2%) contributed most to the modification of AD-risk.

Our approach to map variants to associated genes and to map genes to pathways resulted in a weighted annotation of variants to pathways that allowed for uncertainty in gene as well as pathway assignment, which was not done previously. We note that considering uncertainty in variant-gene as well as gene-pathway assignments is crucial because most genetic variants are in non-coding regions, which makes the closest gene not necessarily the culprit gene, and because different functional annotation-sources often do not overlap. In our variant-pathway mapping, a larger number of annotations (both variant-genes and gene-pathways), generally causes a dilution of the "true" variant effect, reflecting increasing uncertainty in the annotation sources used. This depends on the specific regions, for example, the HLA region carries many genes with large linkage signals, however, all genes in this region are typically annotated with immune response. We point out that the power of the PRSs does not only reflect the effect-size of the variants, but also the number and frequency of the variants that contribute

to the PRSs: due to this, a larger number of very common variants with relatively small effect-size can still have more power (yet small ORs) than a small number of relatively rare variants with high effect-size. The pathway-specific PRS that we proposed in this manuscript can be re-used for the identification of subtypes of AD patients compromised in a specific AD-associated pathway. This is of interest for clinical trials, in order to test responsiveness to compounds in specific subsets of patients. For example, monoclonal antibody targeting *TREM2* receptors could work better in AD patients who have an impaired immune response pathway. Recently, several studies attempted to construct pathway-specific PRS to find heterogeneity in AD patients based on a genetic basis.[28, 29] In line with our findings, *Ahmad et al.* found that genes capturing endocytosis pathway significantly associated with AD and with the conversion to AD.[29] Other studies used less variants [28] or less stringent selection for variants, and did not observe a differential involvement of pathways in AD etiology.[56]

The amyloid cascade hypothesis has been dominating AD-related research in the last two decades. However, treatments targeting amyloid have, so far, not been able to slow or stop disease progression. This has led to an increased interest for the other pathways that are important in AD pathogenesis.[22] Part of the current view of the etiology of AD is that the dysregulation of the immune response is a major causal pathway, and that AD is not only a consequence of β-amyloid metabolism.[57, 58] In addition, previous studies showed that healthy immune and metabolic systems are associated with longer and healthier lifespan.[1, 59] Our results indicate that, excluding *APOE* variants, the effect of immune response and endocytosis on escaping AD is stronger or comparable to the effect on causing AD. This suggests that these pathways might be involved in the maintenance of general cognitive health, as the cognitively healthy centenarians represent the escape of all neurodegenerative diseases until extreme ages. We recently found evidence for this hypothesis in the protective low frequency variant in *PLCG2*, which is involved in the regulation of the immune response.[52] This variant is enriched in cognitively healthy centenarians, and protects against AD as well as frontotemporal dementia and dementia with Lewy bodies. We included this variant in the total PRS as well as in the pathway-PRS for the immune response (variant-pathway mapping was 60%) and endocytosis (variant-pathway mapping was 40%). Regarding endocytosis, this pathway is thought to play a role both in neurons, as part of the β-amyloid metabolism, but also in glia cells, as part of the immune response. Thus, a dysregulation in the interplay between these pathways might lead to an imbalance of immune

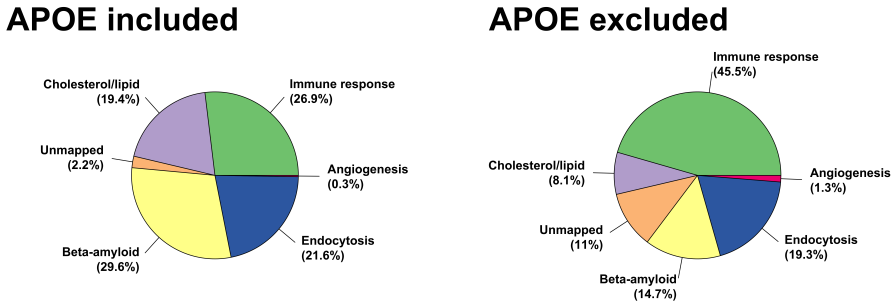


Figure 3.4: **Explained variance of each pathway-specific PRS to polygenic risk of AD** The pie charts represents the explained variance of each pathway-specific PRS to the polygenic risk of AD, including and excluding *APOE* variants. The contributions are calculated according to (*i*) our variant-pathway mapping, (*ii*) the effect size (log of odds ratio) of each variant from literature (Table S1), and (*iii*) variant's frequency in our cohort of middle-aged healthy population subjects. We also considered variants with missing variant-pathway mapping (unmapped pathway).

signaling factors, favoring the engulfment of synapses and AD-associated processes. This, in turn, may contribute to the accumulation of amyloid and tau pathologies.[60, 61, 62, 63]

We assessed the effect of common and low frequency variants on the development and the escape of AD. Therefore, the contributions of rare, causative variants associated with increased AD risk, such as those in APP, *PSEN1, PSEN2, TREM2* and *SORL1* were not considered. Despite the large odds ratios to develop AD associated with carrying such variants, the frequency of these variants in the population is ultra-low, and therefore have a minor effect on the total AD risk in the population.[11, 12] However, future versions of the PRS will most likely include the effect of carrying disease-associated rare variants. This will affect individual PRS scores and the necessity to accordingly adapt the results generated with current PRSs. Compared to the sizes of recent GWAS of AD, we included relatively small sample sizes, particularly with respect to the cognitively healthy centenarians, a very rare phenotype in the population (<0.1%).[4] These sample sizes are however sufficient to study PRSs. The cohorts that we used in this study were not used in any GWAS of AD, therefore we provide independent replication of AD PRS in a homogeneous group of (Dutch) individuals.

We note that, apart from *APOE* variants (for which we stratify the analyses for), none of the other variants have been associated with longevity

or well cognitive functioning in the largest and most recent GWAS.[64, 65] We acknowledge that our variant-pathway mapping reflects the current state of imperfect knowledge at the level of AD-GWAS findings, variant-gene and gene-mechanism relationships. Thus, as new variants, pathways or functional relationships will be identified, the contributions and the pathway-specific PRSs will need to be recalculated. Of note: the study in which the genome-wide significant association with AD of the variant in/near *KANSL1* was originally identified, reported a larger effect size compared to the effect size used in our manuscript, (β=0.31 and β=0.07, respectively), possibly because the original analysis was stratified by *APOE*. We cannot exclude that we underestimated the contribution of *KANSL1* in the analyses. Moreover, since the *KANSL1* variant did not map into one of the analyzed pathways, it was not included in any of the pathway-specific PRS calculations. A limitation, not exclusive to our work, is the highly debated role of *APOE* gene. We mapped the effect of *APOE* to four pathways and we are aware this assignment is relatively arbitrary. We add that *APOE* has well-studied (cardio)vascular properties that are included in our cholesterol and lipid metabolism pathway. The combination of a large effect and unclear pathway assignment makes that pathway-PRS including *APOE* challenging to use. Lastly, we point out that the variance contributions might change in different populations, as it depends on variant frequency and population heterogeneity.

Concluding, with the exclusion of *APOE* variants and based on our functional annotation of variants, the aggregate contribution of the immune response and endocytosis represents more than 60% of the currently known polygenic risk of AD. This indicates that an intervention in these systems may have large potential to prevent AD and potentially other related diseases and highlights the critical need to study (neuro)immune response and endocytosis, next to β-amyloid metabolism.

3.5 Acknowledgements

The following studies and consortia have contributed to this manuscript: Amsterdam dementia cohort (ADC): Research of the Alzheimer center Amsterdam is part of the neurodegeneration research program of Amsterdam Neuroscience (www.amsterdamresearch.org). The Alzheimer Center Amsterdam is supported by Stichting Alzheimer Nederland and Stiching VUmc fonds. The clinical database structure was developed with funding from Stichting Dioraphte. Genotyping of the Dutch case-control samples was performed in the context of EADB (European Alzheimer DNA biobank) funded by the JPco-fuND FP-829-029 (ZonMW projectnumber 733051061). 100-plus study: we are grateful for the collaborative efforts of all participating centenarians and their family members and/or relatives. This work was supported by Stichting Alzheimer Nederland (WE09.2014-03), Stichting Diorapthe, horstingstuit foundation, Memorabel (ZonMW projectnumber 733050814) and Stichting VUmc Fonds. Genotyping of the 100-plus study was performed in the context of EADB (European Alzheimer DNA biobank) funded by the JPco-fuND FP-829-029 (ZonMW projectnumber 733051061). The clinical database structure was developed with funding from Stichting Dioraphte. Longitudinal Aging Study Amsterdam (LASA) is largely supported by a grant from the Netherlands Ministry of Health, Welfare and Sports, Directorate of Long-Term Care. The authors are grateful to all LASA participants, the fieldwork team and all researchers for their ongoing commitment to the study. This work was in part carried out on the Dutch national e-infrastructure with the support of SURF Cooperative. **Conflict of interest**: the authors declare no conflict of interest.

3.6 Full author list and affiliations

Niccolo' Tesi,[1,2,3] Sven J. van der Lee,[1,2] Marc Hulsman,[1,2,3] Iris E. Jansen,[1,4] Najada Stringa,[5] Natasja M. van Schoor,[5] Martijn Huisman,[5] Philip Scheltens,[1] Marcel J.T. Reinders,[3] Wiesje M. van der Flier,[1,5] and Henne Holstege [1,2,3]

[1] Alzheimer Centre, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[2] Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[3] Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

[4] Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, VU, Amsterdam, The Netherlands

[5] Department of Epidemiology and Data Sciences, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

3.7 Supplementary Figures

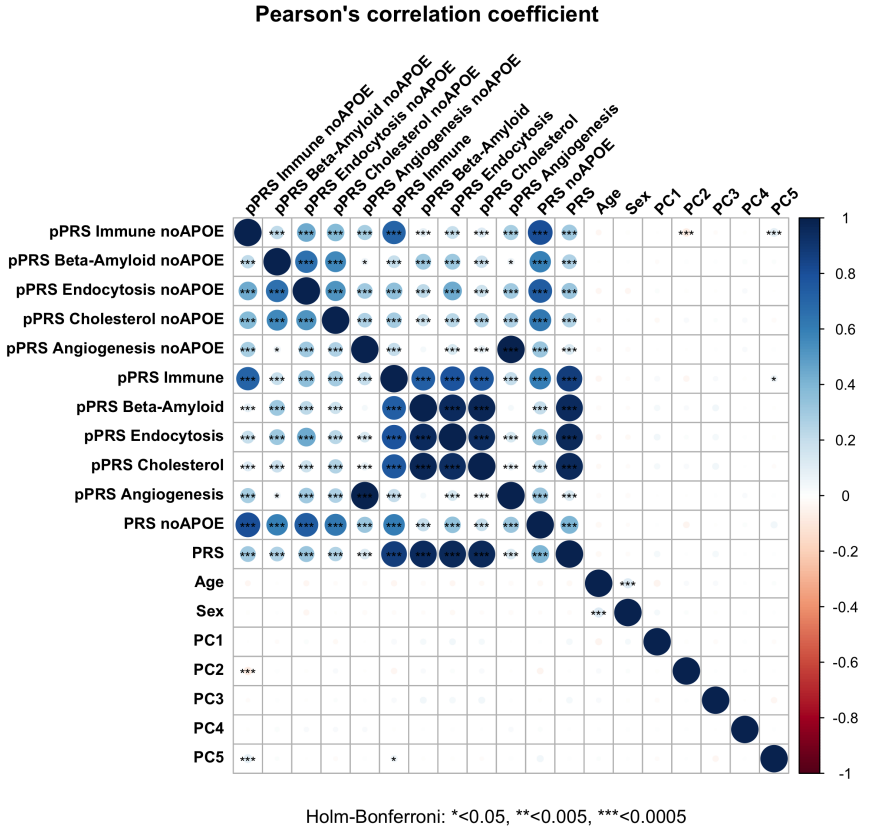


Figure 3.5: **Correlation plot of the (p)PRS and covariates.** The figure shows the correlation between the pathway-PRS (respectively with and without *APOE* variants), the full-PRS (respectively with and without *APOE* variants), ages (ages at study inclusion for controls, ages at diagnosis for AD cases) and the 5 principal components derived from the population stratification analysis. We used Pearson correlation and p-values were adjusted with Holm-Bonferroni correction.

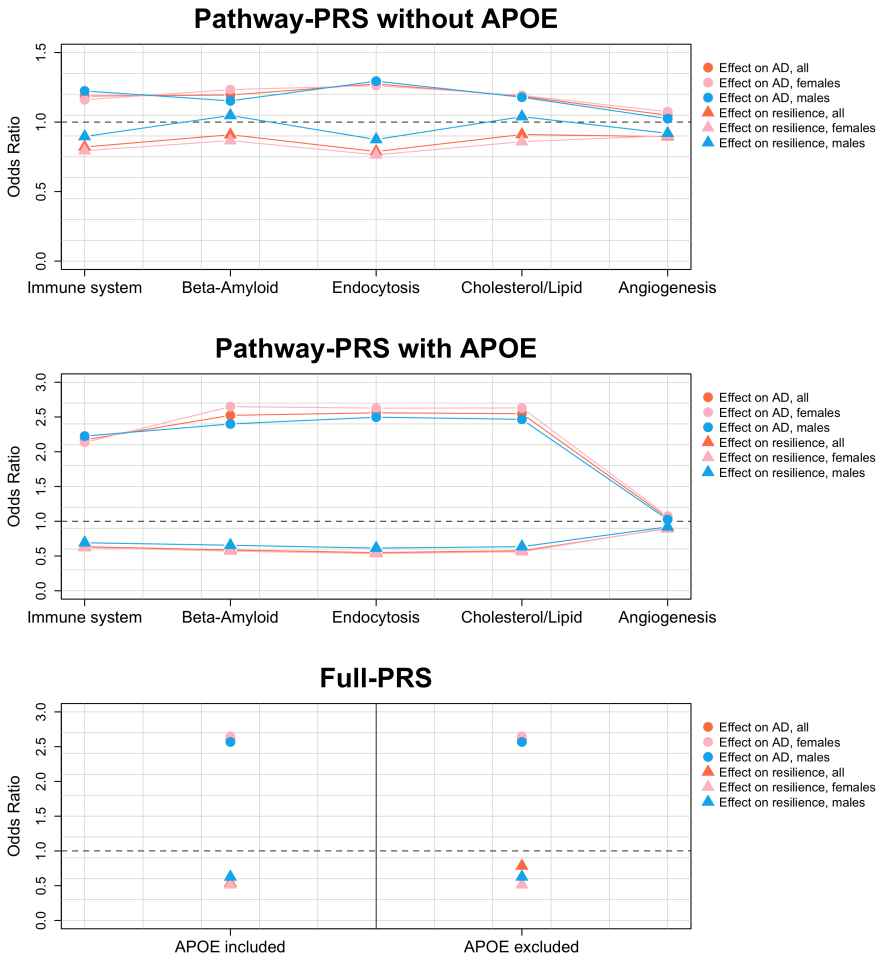


Figure 3.6: **Sex-stratified analysis of the (p)PRS.** The figure shows the sex-stratified analyses in the context of the overall analysis, respectively for the pathway-PRS (including and excluding *APOE* variants) and the Full-PRS (including and excluding *APOE* variants). For each PRS (pathway-PRS or Full- PRS) we report both the odds ratio for AD and those for AD-resilience.

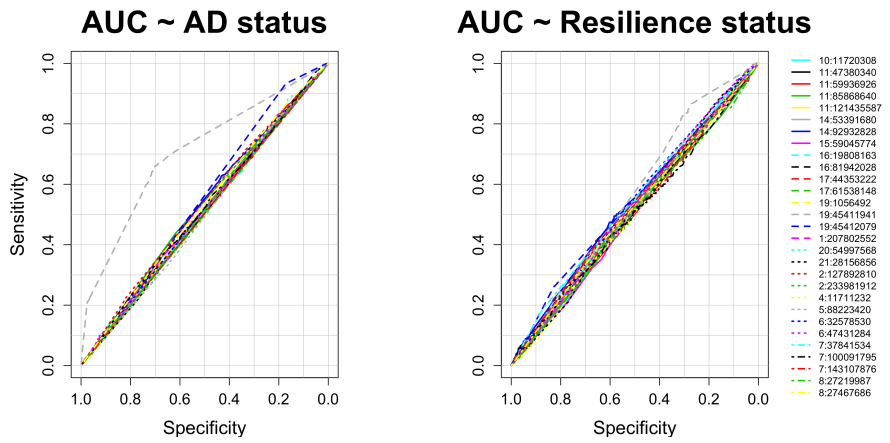


Figure 3.7: **Area under ROC curve for classification of AD or AD-Resilience status, for each variant.** The figure shows the quality of the classification of AD and AD-Resilience status using single-variants.

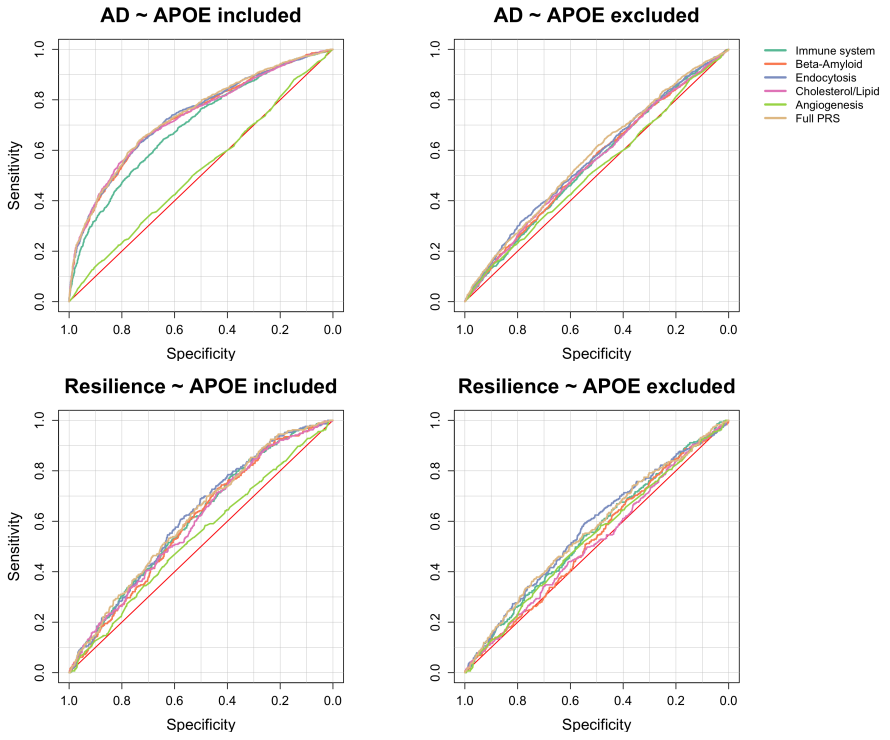
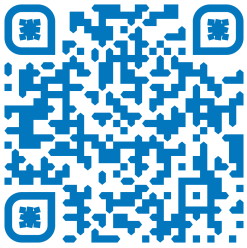


Figure 3.8: **Area under ROC curve for classification of AD and AD- Resilience status, for each PRS and pPRS.** The figure shows the quality of the classification of AD and AD-Resilience status using PRS and pPRS.

3.8 Supplementary Tables

Supplementary Tables can be accessed by scanning the following code or accessing the journal's website here.



References

[1] Linda Partridge, Joris Deelen, and P. Eline Slagboom. "Facing up to the global challenges of ageing". In: *Nature* 561.7721 (Sept. 2018), pp. 45–56. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0457-8.

[2] "2012 Alzheimer's disease facts and figures". In: *Alzheimer's & Dementia* 8.2 (Mar. 2012), pp. 131–168. ISSN: 15525260. DOI: 10.1016/j.jalz.2012.02.001.

[3] María M. Corrada et al. "Dementia incidence continues to increase with age in the oldest old: The 90+ study". In: *Annals of Neurology* 67.1 (Jan. 2010), pp. 114–121. ISSN: 03645134, 15318249. DOI: 10.1002/ana.21915.

[4] Henne Holstege et al. "The 100-plus Study of Dutch cognitively healthy centenarians: rationale, design and cohort description". In: (Apr. 2018). DOI: 10.1101/295287.

[5] Teresa Niccoli and Linda Partridge. "Ageing as a Risk Factor for Disease". In: *Current Biology* 22.17 (Sept. 2012), R741–R752. ISSN: 09609822. DOI: 10.1016/j.cub.2012.07.024.

[6] Angela R. Brooks-Wilson. "Genetics of healthy aging and longevity". In: *Human Genetics* 132.12 (Dec. 2013), pp. 1323–1338. ISSN: 1432-1203. DOI: 10.1007/s00439-013-1342-z.

[7] J. C. Lambert et al. "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease". In: *Nature Genetics* 45.12 (Dec. 2013), pp. 1452–1458. ISSN: 1546-1718. DOI: 10.1038/ng.2802.

[8] Denise Harold et al. "Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease". In: *Nature Genetics* 41.10 (Oct. 2009), pp. 1088–1093. ISSN: 1546-1718. DOI: 10.1038/ng.440.

[9] Sudha Seshadri et al. "Genome-wide analysis of genetic loci associated with Alzheimer disease". In: *JAMA* 303.18 (May 2010), pp. 1832–1840. ISSN: 1538-3598. DOI: 10.1001/jama.2010.574.

[10] Rahul S. Desikan et al. "Polygenic Overlap Between C-Reactive Protein, Plasma Lipids, and Alzheimer Disease". In: *Circulation* 131.23 (June 2015), pp. 2061–2069. ISSN: 1524-4539. DOI: 10.1161/CIRCULATIONAHA.115.015489.

[11] Rebecca Sims et al. "Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease". In: *Nature Genetics* 49.9 (Sept. 2017), pp. 1373–1384. ISSN: 1546-1718. DOI: 10.1038/ng.3916.

[12] Rita Guerreiro et al. "TREM2 variants in Alzheimer's disease". In: *The New England Journal of Medicine* 368.2 (Jan. 2013), pp. 117–127. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1211851.

[13] Thorlakur Jonsson et al. "Variant of TREM2 associated with the risk of Alzheimer's disease". In: *The New England Journal of Medicine* 368.2 (Jan. 2013), pp. 107–116. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1211103.

[14] Paul Hollingworth et al. "Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease". In: *Nature Genetics* 43.5 (May 2011), pp. 429–435. ISSN: 1546-1718. DOI: 10.1038/ng.803.

[15] Adam C. Naj et al. "Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease". In: *Nature Genetics* 43.5 (May 2011), pp. 436–441. ISSN: 1546-1718. DOI: 10.1038/ng.801.

[16] G. Jun et al. "A novel Alzheimer disease locus located near the gene encoding tau protein". In: *Molecular Psychiatry* 21.1 (Jan. 2016), pp. 108–117. ISSN: 1476-5578. DOI: 10.1038/mp.2015.23.

[17] Stacy Steinberg et al. "Loss-of-function variants in ABCA7 confer risk of Alzheimer's disease". In: *Nature Genetics* 47.5 (May 2015), pp. 445–447. ISSN: 1546-1718. DOI: 10.1038/ng.3246.

[18] Riccardo E. Marioni et al. "GWAS on family history of Alzheimer's disease". In: *Translational Psychiatry* 8.1 (Dec. 2018), p. 99. ISSN: 2158-3188. DOI: 10.1038/s41398-018-0150-6.

[19] Alzheimer Disease Genetics Consortium (ADGC), et al. "Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing". In: *Nature Genetics* 51.3 (Mar. 2019), pp. 414–430. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-019-0358-2.

[20] Iris E. Jansen et al. "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk". In: *Nature Genetics* 51.3 (Mar. 2019), pp. 404–413. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-018-0311-9.

[21] Niccolò Tesi et al. "Centenarian controls increase variant effect sizes by an average twofold in an extreme case–extreme control analysis of Alzheimer's disease". In: *European Journal of Human Genetics* (Sept. 2018). ISSN: 1018-4813, 1476-5438. DOI: 10.1038/s41431-018-0273-5.

[22] Caroline Van Cauwenberghe, Christine Van Broeckhoven, and Kristel Sleegers. "The genetic landscape of Alzheimer disease: clinical implications and perspectives". In: *Genetics in Medicine* 18.5 (May 2016), pp. 421–430. ISSN: 1098-3600, 1530-0366. DOI: 10.1038/gim.2015.117.

[23] J. Hardy et al. "Pathways to Alzheimer's disease". In: *Journal of Internal Medicine* 275.3 (Mar. 2014), pp. 296–303. ISSN: 09546820. DOI: 10.1111/joim.12192.

[24] Adam C. Naj, Gerard D. Schellenberg, and for the Alzheimer's Disease Genetics Consortium (ADGC). "Genomic variants, genes, and pathways of Alzheimer's disease: An overview". In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 174.1 (Jan. 2017), pp. 5–26. ISSN: 15524841. DOI: 10.1002/ajmg.b.32499.

[25] Jan Verheijen and Kristel Sleegers. "Understanding Alzheimer Disease at the Interface between Genetics and Transcriptomics". In: *Trends in Genetics* 34.6 (June 2018), pp. 434–447. ISSN: 01689525. DOI: 10.1016/j.tig.2018.02.007.

[26] Rachel E. Bennett et al. "Tau induces blood vessel abnormalities and angiogenesis-related gene expression in P301L transgenic mice and human Alzheimer's disease". In: *Proceedings of the National Academy of Sciences* 115.6 (Feb. 2018), E1289–E1298. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1710329115.

[27] Luigi Yuri Di Marco et al. "Vascular dysfunction in the pathogenesis of Alzheimer's disease — A review of endothelium-mediated mechanisms and ensuing vicious circles". In: *Neurobiology of Disease* 82 (Oct. 2015), pp. 593–606. ISSN: 09699961. DOI: 10.1016/j.nbd.2015.08.014.

[28] Burcu F. Darst et al. "Pathway-Specific Polygenic Risk Scores as Predictors of Amyloid-β Deposition and Cognitive Function in a Sample at Increased Risk for Alzheimer's Disease". In: *Journal of Alzheimer's Disease* 55.2 (Nov. 2016). Ed. by Agustín Ruiz, pp. 473–484. ISSN: 13872877, 18758908. DOI: 10.3233/JAD-160195.

[29] Shahzad Ahmad et al. "Disentangling the biological pathways involved in early features of Alzheimer's disease in the Rotterdam Study". In: *Alzheimer's & Dementia* 14.7 (July 2018), pp. 848–857. ISSN: 15525260. DOI: 10.1016/j.jalz.2018.01.005.

[30] Frank Dudbridge. "Power and Predictive Accuracy of Polygenic Risk Scores". In: *PLoS Genetics* 9.3 (Mar. 2013). Ed. by Naomi R. Wray, e1003348. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003348.

[31] Yun Freudenberg-Hua, Wentian Li, and Peter Davies. "The Role of Genetics in Advancing Precision Medicine for Alzheimer's Disease-A Narrative Review". In: *Frontiers in Medicine* 5 (2018), p. 108. ISSN: 2296-858X. DOI: 10.3389/fmed.2018.00108.

[32] M. Huisman et al. "Cohort Profile: The Longitudinal Aging Study Amsterdam". In: *International Journal of Epidemiology* 40.4 (Aug. 2011), pp. 868–876. ISSN: 0300-5771, 1464-3685. DOI: 10.1093/ije/dyq219.

[33] Emiel O. Hoogendijk et al. "The Longitudinal Aging Study Amsterdam: cohort update 2016 and major findings". In: *European Journal of Epidemiology* 31.9 (Sept. 2016), pp. 927–945. ISSN: 0393-2990, 1573-7284. DOI: 10.1007/s10654-016-0192-0.

[34] Wiesje M. van der Flier and Philip Scheltens. "Amsterdam Dementia Cohort: Performing Research to Optimize Care". In: *Journal of Alzheimer's Disease* 62.3 (Mar. 2018). Ed. by George Perry, Jesus Avila, and Xiongwei Zhu, pp. 1091–1111. ISSN: 13872877, 18758908. DOI: 10.3233/JAD-170850.

[35] Wiesje M. van der Flier et al. "Optimizing patient care and research: the Amsterdam Dementia Cohort". In: *Journal of Alzheimer's disease: JAD* 41.1 (2014), pp. 313–327. ISSN: 1875-8908. DOI: 10.3233/JAD-132306.

[36] Marleen C. Rademaker, Geertje M. de Lange, and Saskia J.M.C. Palmen. "The Netherlands Brain Bank for Psychiatry". In: *Handbook of Clinical Neurology*. Vol. 150. Elsevier, 2018, pp. 3–16. ISBN: 978-0-444-63639-3. DOI: 10.1016/B978-0-444-63639-3.00001-3.

[37] Jean-Charles Lambert et al. "Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease". In: *Nature Genetics* 41.10 (Oct. 2009), pp. 1094–1099. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.439.

[38] W. J. Strittmatter et al. "Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease". In: *Proceedings of the National Academy of Sciences of the United States of America* 90.5 (Mar. 1993), pp. 1977–1981. ISSN: 0027-8424.

[39] E. H. Corder et al. “Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease”. In: *Nature Genetics* 7.2 (June 1994), pp. 180–184. ISSN: 1061-4036. DOI: 10.1038/ng0694-180.

[40] Sayantan Das et al. “Next-generation genotype imputation service and methods”. In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1284–1287. ISSN: 1546-1718. DOI: 10.1038/ng.3656.

[41] Shane McCarthy et al. “A reference panel of 64,976 haplotypes for genotype imputation”. In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1279–1283. ISSN: 1546-1718. DOI: 10.1038/ng.3643.

[42] 1000 Genomes Project Consortium et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (Oct. 2015), pp. 68–74. ISSN: 1476-4687. DOI: 10.1038/nature15393.

[43] Wilfred A Jefferies et al. “Adjusting the compass: new insights into the role of angiogenesis in Alzheimer’s disease”. In: *Alzheimer’s Research & Therapy* 5.6 (2013), p. 64. ISSN: 1758-9193. DOI: 10.1186/alzrt230.

[44] Anthony H Vagnucci and William W Li. “Alzheimer’s disease and angiogenesis”. In: *The Lancet* 361.9357 (Feb. 2003), pp. 605–608. ISSN: 01406736. DOI: 10.1016/S0140-6736(03)12521-4.

[45] Kyoko Watanabe et al. “Functional mapping and annotation of genetic associations with FUMA”. In: *Nature Communications* 8.1 (Dec. 2017), p. 1826. ISSN: 2041-1723. DOI: 10.1038/s41467-017-01261-5.

[46] The Gene Ontology Consortium. “Expansion of the Gene Ontology knowledgebase and resources”. In: *Nucleic Acids Research* 45 (D1 Jan. 2017), pp. D331–D338. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkw1108.

[47] M. Ashburner et al. “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium”. In: *Nature Genetics* 25.1 (May 2000), pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556.

[48] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists”. In: *Nucleic Acids Research* 37.1 (Jan. 2009), pp. 1–13. ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gkn923.

[49] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources”. In: *Nature Protocols* 4.1 (Jan. 2009), pp. 44–57. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/nprot.2008.211.

[50] Valentina Escott-Price et al. “Polygenic risk score analysis of pathologically confirmed Alzheimer disease: PRS Analysis of AD”. In: *Annals of Neurology* 82.2 (Aug. 2017), pp. 311–314. ISSN: 03645134. DOI: 10.1002/ana.24999.

[51] Valentina Escott-Price et al. “Polygenic score prediction captures nearly all common genetic risk for Alzheimer’s disease”. In: *Neurobiology of Aging* 49 (Jan. 2017), 214.e7–214.e11. ISSN: 01974580. DOI: 10.1016/j.neurobiolaging.2016.07.018.

[52] DESGESCO (Dementia Genetics Spanish Consortium), EADB (Alzheimer Disease European DNA biobank) et al. “A nonsynonymous mutation in PLCG2 reduces the risk of Alzheimer’s disease, dementia

with Lewy bodies and frontotemporal dementia, and increases the likelihood of longevity". In: *Acta Neuropathologica* (May 2019). ISSN: 0001-6322, 1432-0533. DOI: 10.1007/s00401-019-02026-8.

[53] for the International Genomics of Alzheimer's Project et al. "Evaluation of a Genetic Risk Score to Improve Risk Prediction for Alzheimer's Disease". In: *Journal of Alzheimer's Disease* 53.3 (Aug. 2016). Ed. by Anette Hall, pp. 921–932. ISSN: 13872877, 18758908. DOI: 10.3233/JAD-150749.

[54] Sultan Chaudhury et al. "Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment". In: *Translational Psychiatry* 9.1 (Dec. 2019), p. 154. ISSN: 2158-3188. DOI: 10.1038/s41398-019-0485-7.

[55] Sven J van der Lee et al. "The effect of APOE and other common genetic variants on the onset of Alzheimer's disease and dementia: a community-based cohort study". In: *The Lancet Neurology* 17.5 (May 2018), pp. 434–444. ISSN: 14744422. DOI: 10.1016/S1474-4422(18)30053-X.

[56] Ganna Leonenko et al. "Genetic risk for Alzheimer's disease is distinct from genetic risk for amyloid deposition". In: *Annals of Neurology* (June 2019). ISSN: 0364-5134, 1531-8249. DOI: 10.1002/ana.25530.

[57] R. M. Ransohoff. "How neuroinflammation contributes to neurodegeneration". In: *Science* 353.6301 (Aug. 2016), pp. 777–783. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aag2590.

[58] Michael T Heneka et al. "Neuroinflammation in Alzheimer's disease". In: *The Lancet Neurology* 14.4 (Apr. 2015), pp. 388–405. ISSN: 14744422. DOI: 10.1016/S1474-4422(15)70016-5.

[59] Peter K. Joshi et al. "Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity". In: *Nature Communications* 8.1 (Dec. 2017). ISSN: 2041-1723. DOI: 10.1038/s41467-017-00934-5.

[60] Karpagam Srinivasan et al. "Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses". In: *Nature Communications* 7.1 (Dec. 2016). ISSN: 2041-1723. DOI: 10.1038/ncomms11295.

[61] Marie Orre et al. "Isolation of glia from Alzheimer's mice reveals inflammation and dysfunction". In: *Neurobiology of Aging* 35.12 (Dec. 2014), pp. 2746–2760. ISSN: 01974580. DOI: 10.1016/j.neurobiolaging.2014.06.004.

[62] Yaming Wang et al. "TREM2-mediated early microglial response limits diffusion and toxicity of amyloid plaques". In: *The Journal of Experimental Medicine* 213.5 (May 2016), pp. 667–675. ISSN: 0022-1007, 1540-9538. DOI: 10.1084/jem.20151948.

[63] David V. Hansen, Jesse E. Hanson, and Morgan Sheng. "Microglia in Alzheimer's disease". In: *The Journal of Cell Biology* 217.2 (Feb. 2018), pp. 459–472. ISSN: 0021-9525, 1540-8140. DOI: 10.1083/jcb.201709069.

[64] Paul RHJ Timmers et al. "Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances". In: *eLife* 8 (Jan. 2019). ISSN: 2050-084X. DOI: 10.7554/eLife.39856.

[65] Gail Davies et al. “Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function”. In: *Nature Communications* 9.1 (Dec. 2018). ISSN: 2041-1723. DOI: 10.1038/s41467-018-04362-x.

4. The Alzheimer-Longevity axis

The effect of Alzheimer's disease-associated genetic variants on longevity

Niccolo' Tesi, Marc Hulsman, Sven J. van der Lee, Iris E. Jansen, Najada Stringa, Natasja M. van Schoor, Martijn Huisman, Philip Scheltens, Wiesje M. van der Flier, Marcel J.T. Reinders, and Henne Holstege

This chapter was submitted
https://doi.org/10.1101/2021.02.02.21250991

Abstract

Human longevity is influenced by the genetic risk of age-related diseases. As Alzheimer's disease (AD) represents a common condition at old age, an interplay between genetic factors affecting AD and longevity is expected. We explored this interplay by studying the prevalence of AD-associated single-nucleotide-polymorphisms (SNPs) in cognitively healthy centenarians, and replicated findings in a parental-longevity GWAS. We found that 28/38 SNPs that increased AD-risk also associated with lower odds of longevity. For each SNP, we express the imbalance between AD- and longevity-risk as an effect-size distribution. Based on these distributions, we grouped the SNPs in three groups: 17 SNPs increased AD-risk more than they decreased longevity-risk, and were enriched for β-amyloid metabolism and immune signaling; 11 variants reported a larger longevity-effect compared to their AD-effect, were enriched for endocytosis/immune-signaling, and were previously associated with other age-related diseases. Unexpectedly, 10 variants associated with an increased risk of AD and higher odds of longevity. Altogether, we show that different AD-associated SNPs have different effects on longevity, including SNPs that may confer general neuro-protective functions against AD and other age-related diseases.

4.1 Introduction

The human lifespan is determined by a beneficial combination of environmental and genetic factors.[1, 2] Long-lived individuals tend to cluster in families, suggesting that the role of the genetic factors is considerable,[3, 4] however, the research of genetic variants that influence human lifespan has yielded contrasting results: only the longevity-association of the *APOE* alleles and few additional variants consistently replicated across studies (*CDKN2B, ABO*).[5, 6] While the replication rate in independent studies is low, a large collection of genetic variants has been associated with longevity through genome-wide association studies (GWAS) in different studies and populations.[5, 6] The majority of these variants was previously identified to associate with other age-related conditions, including cardiovascular disease, autoimmune and neurological disorders, suggesting that the genetics underlying human longevity depends on a lower risk for several age-related diseases.[5, 6, 2] Of all age-related diseases, late-onset Alzheimer's Disease (AD) is the most common type of dementia and one of the most prevalent causes of death at old age.[7] The largest risk factor for AD is aging: at 100 years of age, the disease's incidence is about 40% per year.[8] Genetic factors play a significant role in AD as heritability was estimated to be 60-80%:[9] the strongest common genetic risk factor for AD is the *APOE* ϵ4 allele, and large collaborative GWAS have identified 41 additional common variants associated with a slight modification of the risk of AD.[10, 11, 12, 13] Despite high incidence rates of AD at very old ages, AD is not an inevitable consequence of aging, as demonstrated by individuals who surpass the age of 100 years with high levels of cognitive functioning.[14] As AD-associated variants increase the risk of AD, leading to earlier death, a negative effect on longevity for these variants should be expected. However, apart from *APOE* alleles, genetic variants that influence the risk of AD were not found to affect the human lifespan in previous GWAS. In fact, often we assume that AD-associated variants affect AD only, but this may still not hold true. For example, at the molecular level, there may be other age-related traits that share (part of) the biological pathways underlying AD. Nevertheless, for an AD-associated variant that affects AD only, the relative effect on longevity should be proportional to the corresponding effect on AD, albeit in a different direction. This means that if a variant increases the risk of AD 2-fold, then carriers will have twice as much AD-related mortality as non-carriers, and as a consequence, they will have twice as little chance to age into a cognitively healthy centenarian. However, in case a genetic variant is protective against

multiple conditions, it might be expected that the overall effect on longevity results larger than the absolute effect on AD alone.

We have previously shown that cognitively healthy centenarians are depleted with genetic variants that increased the risk of AD compared to a general population. Yet, the extent of depletion was variant specific, suggesting that a subset of AD-variants may be specifically beneficial to reach extremely old ages in good cognitive health.[15, 16] In addition, the extent to which these variants affect other age-related diseases is mostly unknown.[17] Using the notion of effect-size proportionality, we set out to investigate the relationship between AD- and longevity- risk for genetic variants associated with AD.

4.2 Methods

4.2.1 Populations and selection of genetic variants

We included *N*=358 centenarians from the 100-plus Study cohort, which comprises Dutch-speaking individuals aged 100 years or older who self-report to be cognitively healthy, which is confirmed by a proxy.[14] As population controls, we used population-matched, cognitively healthy individuals from five studies: (*i*) the Longitudinal Aging Study of Amsterdam (LASA, *N*=1,779),[18, 19] (*ii*) the memory clinic of the Alzheimer center Amsterdam and SCIENCe project (*N*=1,206),[20, 21] (*iii*) the Netherlands Brain Bank (*N*=40),[22] (*iv*) the twin study of Amsterdam (*N*=201)[23] and (*v*) the 100-plus Study (partners of centenarian's children, *N*=86).[14] See *Supplementary Methods: Populations* for a detailed description of these cohorts. Throughout the manuscript, we will refer to the union of the individuals from these five studies as population subjects. The Medical Ethics Committee of the Amsterdam UMC (METC) approved all studies. All participants and/or their legal representatives provided written informed consent for participation in clinical and genetic studies. Genetic variants in our populations were determined by standard genotyping and imputation methods. All samples were genotyped using the same commercial kit. After establishing quality control of the genetic data (see *Supplementary Methods: Quality control*), 2,905 population subjects and 343 cognitively healthy centenarians were left for the analyses (Table 4.1). We then selected 41 variants representing the current genetic landscape of AD (Table S1).[13] We restricted our analysis to high-quality variants with a minor allele frequency >1% in our cohorts, which led to the exclusion of 3/41 variants (rare variants in the *TREM2* gene *rs143332484* and *rs75932628* and *ABI3* gene *rs616338*), leaving 38 variants for the analyses.

4.2.2 AD and longevity variant effect sizes

We first retrieved the effect-size on AD (E^k_{AD}) for each AD variant, k, from one of the largestGWAS of AD.[13] To estimate a confidence interval, we sampled (S=10000) from the published effect-sizes (log of odds ratios) and their respective standard errors. To calculate the effect-size on longevity (E^k_{LGV}) for the same variants, we used a logistic regression model with cognitive healthy centenarians as cases and population subjects as controls while adjusting for population stratification (PC 1-5). The number of principal components to include as covariates was arbitrarily chosen; however, as all individuals were population-matched, we expected these components to

correct all major population effects. The resulting p-values were corrected for multiple testing (False Discovery Rate, FDR). To calculate the confidence interval, we repeated this bootstrapping procedure (B=10,000) of the data. For convenience, variant effect-sizes on AD and longevity were calculated with respect to the allele that increases the risk of AD, such that $E^k_{AD} > 0$. Given a variant k, with a relative effect-size on AD (E^k_{AD}) and on longevity (E^k_{LGV}), we defined that the variant has an expected direction if the variant increases the risk of AD, *i.e.* $E^k_{AD} > 0$, and at the same time decreases the risk of longevity, *i.e.* $E^k_{LGV} < 0$. Inversely, we define that the longevity effect has an unexpected direction if the allele that increased AD risk also increased the risk of longevity, *i.e.* $E^k_{AD} > 0$ and $E^k_{LGV} > 0$. The probability of observing an expected direction was considered a Bernoulli variable with p=0.5 (*i.e.* equal chance of having an expected/unexpected direction), thus the number of variants with an expected direction follows a binomial distribution.

4.2.3 Imbalance of variant effect direction

We represented each variant as a data point whose coordinates were defined by the variant's effect on AD (E^k_{AD}, on the y-axis) and its effect on longevity (E^k_{LGV}, on the x-axis). See Figure 4.4 for an example. For each variant, we then calculated the normalized angle, α_k, of the vector representing the data point with the x-axis:

$$\alpha_k = \frac{atan2(E^k_{AD}, E^k_{LGV})}{\pi/2} + 1 \tag{4.1}$$

where $\alpha_k \in [-1;1]$. This normalized angle relates to the imbalance between the risk of AD and the risk of longevity. That is, for $\alpha_k < 0$ the variant has an expected direction, while for $\alpha_k > 0$ the variant has an unexpected direction. As the effect-sizes are sample estimates, we subsequently took their confidence interval into account to create, for each variant, a distribution of the imbalance in the effect direction (*IED*). Hereto, we assumed a Gaussian density for both E^k_{AD} and E^k_{LGV}, centered around $\overline{E}^k_{AD}$ and $\overline{E}^k_{LGV}$ and with a variance equal to the estimated confidence interval for both effect sizes, respectively. We sampled 10,000 times from these distributions and calculated the corresponding imbalance (α_k), to get a (non-Gaussian) distribution of the *IED* for that variant, IED_k. To group variants with similar patterns of their *IED* distributions, we ordered the *IED* by their median value ($\tilde{IED}_k$), and defined a group of variants in which the effect sizes were in the expected direction ($\tilde{IED}_k \leq 0$), which we subsequently split in those that have (*i*) a

larger effect on longevity as compared to the effect on AD ($I\tilde{E}D_k \leq -1/2$, longevity-group), and those that have (*ii*) a larger effect on AD as compared to the effect on longevity ($-1/2 < I\tilde{E}D_k \leq 0$, AD-group). We defined a third group of variants that have an effect in the unexpected direction ($I\tilde{E}D_k > 0$, Unex-group). These cut-off calues were not arbitrarily chosen, instead, they represent the point at which the effect on AD equals the (negative) effect on longevity ($IED_k = -1/2$) and the point at which no effect on longevity is observed ($IED_K = 0$).

4.2.4 Replication of findings in large GWAS cohorts

To find additional evidence for our findings, we inspected the association statistics of the 38 AD-associated variants in the largest GWAS on parental longevity.[6] Briefly, in this study offspring's genotypes were used to model parental age at death. In this dataset, we looked at the significance of association with longevity for the 38 variants (*p*-values were corrected with FDR) and their direction of effect. Finally, we tested the consistency in the expected/unexpected directions between our study and the GWAS on parental longevity using binomial tests. We did not use a case-control GWAS of longevity as the most recent included our cohort, thus the resulting associations would be biased.[5]

4.2.5 Linking variants with functional clusters

To investigate each variant's functional consequences, we calculated the variant-pathway mapping, which indicates the degree of involvement of each genetic variant in AD-associated pathways (Figure 4.5). See Supplementary Methods: variant-pathway mapping, for a detailed explanation of our approach. Briefly, the variant-pathway mapping depends on (*i*) the number of genes each variant was associated with and (*ii*) the biological pathways each gene was associated with. We calculated the variant-pathway mapping for all 38 AD-associated variants. Finally, we compared the variant-pathway mapping within each group of variants defined based on the *IEDs* (Longevity-, AD- and Unex-groups) using Wilcoxon sum rank tests and correcting *p*-values using FDR: this was indicative of whether a group of variants was enriched for a specific functional cluster (Figure 4.5).

4.2.6 Cell-type annotation at the level of each cluster

To further explore the biological basis of the different groups of variants (Longevity-, AD- and Unex-groups), we calculated the degree of enrichment

of each group for specific brain cell-types (see Supplementary Methods: cell-type annotation, for a detailed description). This annotation depends on the number of genes each variant was associated with, and the expression of these genes in the different brain cell-types, *i.e.* astrocytes, oligodendrocytes, microglia, endothelial cells, and neurons. We finally compared the cell-specific annotations within each group of variants (Longevity-, AD- and Unex-groups) using Wilcoxon sum rank tests and correcting p-values using FDR, which indicated whether a group of variants was enriched for specific brain cell-types (Figure 4.5).

4.2.7 Implementation

Quality control of genotype data, population stratification analysis and relatedness analysis were performed with PLINK (v2.0 and v1.9). All subsequent analyses were performed with R (v3.6.3), Bash, and Python (v3.6) scripts. All scripts are freely available at `https://github.com/TesiNicco/Disentangle_AD_Age`.[24] Variant-gene annotation and gene-set enrichment analyses were performed through the web-server that is freely accessible at `https://snpxplorer.net`.

Table 4.1: **Population characteristics**

	Population controls	Cognitively Healthy Centenarians
Individuals	2,905	343
Females (%)	1400 (48.2%)	246 (71.7%)
Age (SD)[a]	68.3 (11.5)	101.4 (1.8)
ApoE $\epsilon 4$ (%)	1012 (17.38)	48 (7.15)
ApoE $\epsilon 2$ (%)	523 (9.00)	91 (13.26)

[a], Age at study inclusion; SD, standard deviation; *ApoE*, Apolipoprotein E allele count for the $\epsilon 4$ and $\epsilon 2$, and relative allele frequency in population controls and cognitively healthy centenarians. References to the cohorts reported in this table are: [21, 19, 14, 22, 23, 20]

4.3 Results

4.3.1 AD-associated variants also associate with longevity

We explored the association with longevity of 38 genetic variants previously associated with AD from GWAS (Table S1). We tested these variants in 343 centenarians who self-reported to be cognitively healthy (mean age at inclusion 101.4±1.3, 74.7% females), as opposed to 2,905 population subjects (mean age at inclusion 68.3±11.5, 50.7% females) (Table 4.1). We found a significant association with longevity for two variants after multiple testing correction (FDR<5%, variants in the *APOE* gene; *rs429358* and *rs7412*, Table S2). We compared the direction of effect on longevity with that on AD as found in literature: of the 38 variants, 28 showed an association in the expected direction, *i.e.* alleles that increased AD risk were associated with lower odds of longevity, and this was significantly more than expected by chance (p=0.005 including *APOE* variants, p=0.01 excluding *APOE* variants, see section 4.2).

4.3.2 Distributions of the imbalance in the effect direction (*IED*)

To study the relationship between the effect on AD and longevity for all 38 AD-associated variants in more detail, we created distributions of the imbalance in the variant effect direction (*IED*): Figure 4.1. The *IED* of a variant indicates (*i*) whether the effects on AD and longevity are in the expected or unexpected direction, and (*ii*) how the effetcs on AD and longevity relate to each other. For example, the two variants *rs7412* and *rs429358* in *APOE* gene significantly associated with longevity in the expected direction and

thus had tight confidence intervals. The resulting *IED* relied completely in the expected direction side (Figure 4.1). In addition, the effect on AD was larger than that on longevity, causing the *IED* to slightly skew towards the AD-side (Figure 4.1). However, as the association of a variant with longevity became less strong (thus with larger confidence intervals) or was in the unexpected direction, the fraction of data points in the unexpected direction increased. For example, for the intergenic variant *rs6733839* close to *BIN1* gene, we observed a larger effect on AD compared to cognitively healthy aging ($E_{AD}^{BIN1} = 0.17$, $SE = 0.01$ and $E_{LGV}^{BIN1} = -0.14$, $SE = 0.08$, $p = 0.09$), yet in the expected direction. The resulting *IED* is skewed towards the AD side (Figure 4.1), and, due to large confidence intervals on longevity, we observed data points in the unexpected direction. Finally, variant *rs593742* near *ADAM10* gene ($E_{AD}^{ADAM10} = 0.08$, $SE = 0.01$ and $E_{LGV}^{ADAM10} = 0.06$, $SE = 0.09$, $p = 0.49$) associated with higher odds of both AD and longevity (unexpected direction of effect), with a resulting *IED* largely on the unexpected side with fewer data points on the expected direction (due to large confidence intervals).

4.3.3 Grouping variants based on *IED* distributions

Based on the median value of each *IED* distributions, $\tilde{IED}_k$, we grouped the variants into (*i*) a Longevity-group (variants with a $\tilde{IED}_k$ skewed towards the longevity-end of the spectrum), (*ii*) an AD-group (variants with a $\tilde{IED}_k$ skewed towards the AD-end of the spectrum), and (*iii*) an Unex-group (variants with a $\tilde{IED}_k$ in the unexpected direction). The AD-group included 17 variants (in/near genes *APOE* (1), *APOE* (2), *SCIMP*, *PLCG2* (1), *MS4A6A*, *BIN1*, *PILRA*, *APP*, *PLCG2* (2), *CR1*, *SLC24A4*, *TREML2*, *ACE*, *APH1B*, *FERMT2*, *PICALM*, *CD33*) and the longevity-group included 11 variants (in or near genes *SHARPIN* (1), *SHARPIN* (2), *HS3ST1, EPAH1, IQCK, PRKD3, CD2AP, PLCG2* (3), *SPI1, HLA, EDHDC3*), such that the effect of 28/38 (74%) of all variants was in the expected direction. The effect of 10 variants was in the unexpected direction, the Unex-group: (*PTK2B, CLU, KANSL1, INPP5D, ABCA7, CHRNE, SORL1, IL34, ADAM10, CASS4*) (Figure 4.1).

4.3.4 AD-associated variants in large GWAS of longevity

To find additional evidence for longevity associations, we inspected the AD-associated variants' effect in the largest GWAS on parental longevity.[6] Of the 38 AD-associated variants, association statistics were available for 34 of the variants (missing from Longevity-group: *PLCG2* (3), *SPI1*; missing from

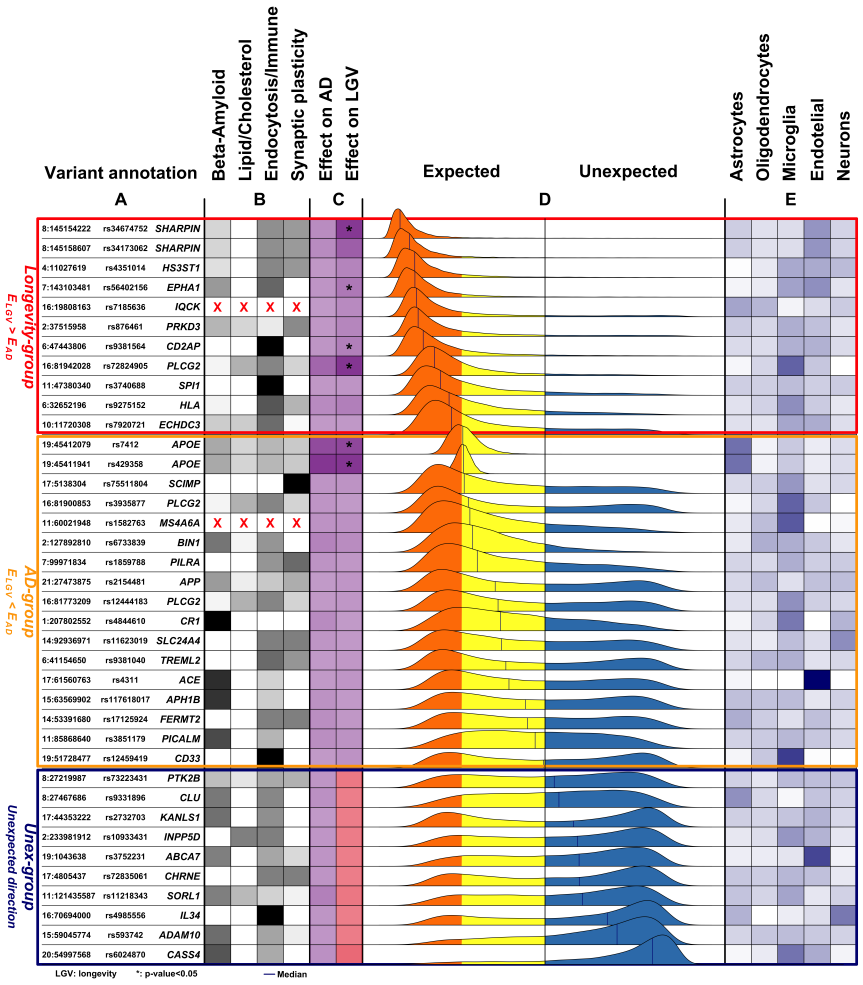


Figure 4.1: **Overview of the 38 genetic variants associated with Alzheimer's disease. A.** Genomic coordinates (with respect to GRCh37), variant identifier, and closest gene. **B.** The variant-pathway mapping score for the four functional clusters (darker colors represent stronger associations). Variants reporting red crosses could not be annotated as no biological processes were found to be associated with the related genes. **C.** The effect size on AD (from literature) and the observed effect size on longevity (LGV) for each variant (darker colors indicate stronger effect). The same color indicates expected direction (*i.e.* increased risk of AD and decreased chance of longevity), while different colors indicates unexpected direction. For the longevity effects, we also annotate variants with a significant association (*, unadjusted p-value<0.05). **D.** The distribution of the imbalance direction of effect (*IED*) of AD-risk and longevity (see section 4.2 for details). The Longevity-, AD- and Unex-groups were derived based on the median value of the *IED* (blue vertical line). **E.** Average expression of the genes associated with the variants in five different brain cell-types (the darker, the higher the expression).

Unex-group: *KANSL1, INPP5D*). Overall, 21/26 (81%) of the variants in the expected direction in our study (of which 6/9 variants in Longevity- and 15/17 variants in the AD-group), were also in the expected direction in the independent parental longevity dataset. Variants in the expected direction in the first analysis are significantly more likely to be in the expected direction in the replication analysis than in the unexpected direction (p=0.01, based on a binomial test, Figure 4.3). Six AD-associated variants reached significance in the parental-longevity GWAS after correcting for multiple comparisons (FDR<5%): variants in the *APOE* gene (*rs429358* and *rs7412*) and variants in/near *PRKD3* (*rs8764613*), *CD2AP* (*rs9381564*), *APH1B* (*rs117618017*) and *BIN1* (*rs6733839*). Of these, variants in/near *PRKD3* and *CD2AP* belonged to the Longevity-group in our analysis. Conversely, only 2/8 (25%) variants that we observed in the unexpected direction in our study were also in the unexpected direction in the parental-longevity GWAS, such that these variants were not more likely to be in the unexpected direction (p=0.29, based on a binomial test, Figure 4.2).

4.3.5 Functional characterization of variants

The 38 AD-associated variants included coding variants (N=10), intronic variants (N=20), and intergenic variants (N=8) (Table S3). 12/28 of the intronic/intergenic variants had eQTL associations. In total, the 38 variants mapped to 68 unique genes, with most variants mapping to one gene (N=21) and fewer mapping to 2 genes (N=10), 3 genes (N=2), 4 genes (N=1), 5 genes (N=2), 6 and 7 genes (N=1, respectively) (Figure 4.6 and Table S3). We performed gene-set enrichment analysis using a sampling-based approach to explore the biological processes enriched in the 68 genes associated with AD-variants (see section 4.2 and Figure 4.5). We found 115 significantly enriched biological processes after correction for multiple tests (FDR<5%, Table S4). After clustering these terms based on their semantic similarity, we found four main clusters of biological processes: (*i*) β-amyloid metabolism, (*ii*) lipid/cholesterol metabolism, (*iii*) endocytosis/immune signaling and (*iv*) synaptic plasticity (Figure 4.1, **??** and Table S5). Next, we calculated the variant-pathway mapping score (see *Methods* and Figure 4.5), which indicates how well a variant is associated with each of the 4 functional clusters. In total, we calculated the variant-pathway mapping for 30 variants; we imputed the annotation of 6 variants (Table S5), while 2 variants could not be annotated (variants *rs7185636* and *rs1582763* in/near *IQCK* and *MS4A6A* genes), because the associated genes were not annotated with any biological process function

(Table S5). Finally, we tested whether the Longevity-, AD- and Unex-groups were enriched for specific functional clusters by comparing the distribution of variant-pathway mapping within each group (see section 4.2, Figure 4.3, and Figure 4.5). The Longevity-group was significantly (FDR<10%) enriched for the endocytosis/immune signaling functional cluster; the AD-group for the endocytosis/immune signaling, β-amyloid metabolism and to a smaller extent for the synaptic plasticity functional clusters; the Unex-group was mainly enriched for the endocytosis and β-amyloid metabolism functional clusters.

4.3.6 Expression of AD-associated genes in brain cell-types

We explored whether specific brain cell types, *i.e.* astrocytes, oligodendrocytes, microglia, endothelial cells and neurons, were enriched within each group of variants (see section 4.2 Table S5, and Table S6). Figure 4.1 shows the collapsed cell-type expression for all 38 AD-associated variants. We then tested the enrichment for cell-type expression within the Longevity-, AD- and Unex-groups. The Longevity-group was significantly enriched for myeloid and endothelial cells, the AD-group for myeloid cells, while the Unex-group was significantly enriched for endothelial cells (FDR<10%, Figure 4.3).

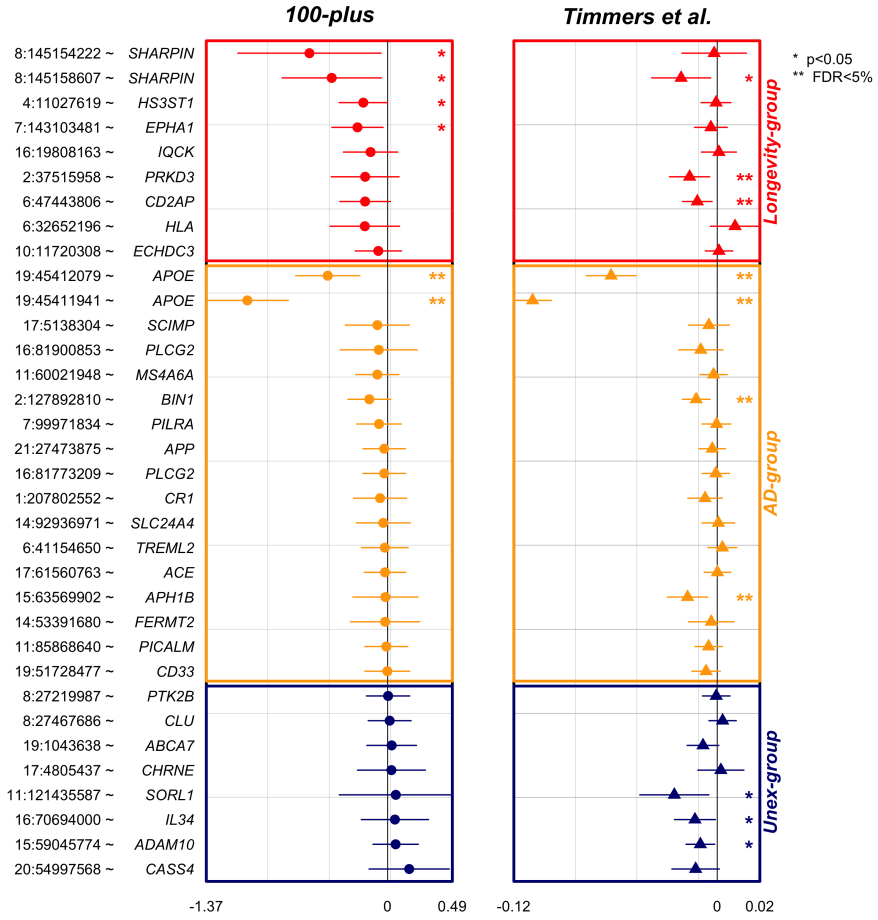


Figure 4.2: **Forest plot of association statistics of AD-variants in our study and the largest GWAS of parental longevity.** The plot shows the association of AD-variants in our study and the largest by-proxy GWAS on parental longevity.[6] The association statistics of 34/38 variants were available from publicly available summary statistics of *Timmers et al.* study. [6] Plotted effect-sizes are with respect to the AD-risk increasing allele. Thus, an expected direction of effect is shown for variants with a negative estimate. Nominally significant associations with AD (p<0.05) are annotated with an asterisk (*), and significant associations after FDR correction are annotated with two asterisks (**).

4.4 Discussion

4.4.1 Summary of the findings

We studied the effect on longevity of 38 genetic variants previously associated with AD through GWAS.[13] We found that a majority of 74% of the alleles that increase the risk of AD is associated with lower odds of becoming a centenarian (expected direction). Overall, most variants (*N*=17) had a larger effect on AD than on longevity: these variants were associated with β-amyloid metabolism and endocytosis/immune signaling, and were primarily expressed in microglia. A subset of variants (*N*=11) had a larger effect on longevity than their effect on AD. These variants were associated mostly with endocytosis and immune signaling, and they were expressed in microglia and endothelial cells. These variant-effects were confirmed for 81% of the alleles in an independent dataset, the largest GWAS on parental longevity. In contrast, 26% of the variants increased both the risk of developing AD and the risk of becoming a centenarian (*N*=10), (unexpected direction). These unexpected effects could only be replicated for 2 of the variants in the independent dataset, suggesting that the expected effects were more robust across studies than the unexpected effects. Together, our findings suggest that a subset of variants associated with AD-risk may also affect longevity, for example through their effect on other age-related diseases.

4.4.2 AD-associated variants and their effect on healthy aging

A single study previously explored the extent to which 10 AD-associated variants affect longevity: apart from *APOE* locus, none of the other 10 tested AD-associated variants significantly associated with longevity.[25] In addition to *APOE*, four variants showed a negative effect on longevity while increasing AD-risk (in/near *ABCA7, EPHA1, CD2AP*, and *CLU*). In agreement with these findings, we also found that only the *APOE* variants significantly associated with longevity, and variants in/near *EPHA1* and *CD2AP* belong to the Longevity-group. However, in our study, we found that most alleles associated with an increased risk of AD associated with a decreased chance of longevity. The inability to observe such an inverse relationship between variant effects on AD and longevity in the previous study may be explained by the relatively small sample sizes, combined with a low number of (well-established) AD variants analyzed (*N*=10). In our study, groups sizes were also relatively small, but the centenarians had a relatively high level of cognitive health, which might have contributed to an increased effect size of AD-associated genetic variants in our comparison.[15, 16]

4.4.3 Different trajectories of effect of AD-associated variants on longevity

Variants with a larger effect on AD than longevity
For most variants with effects in the expected direction, the risk-increasing effect on AD was more extensive than the negative effect on survival/longevity. These variants, which include both *APOE* alleles, might negatively affect lifespan because carriers are removed from the population with increasing age due to AD-associated mortality. For the *APOE* variants specifically, the distribution of the imbalance in the effect directions suggests a nearly similar proportion of the increased risk of AD and decreased risk of longevity for both *APOE* variants $I\tilde{E}D_k \approx 1/2$. This explains why multiple previous studies have associated *APOE* variants with longevity. In our cohort of centenarians, the frequency of the deleterious $\epsilon 4$ allele is half of that of the population controls (8% *vs.* 16%, respectively). In comparison, the frequency of the protective $\epsilon 2$ allele is nearly two-fold increased (16% *vs.* 9%).[15] Note, however, that inclusion criteria of the centenarian cohort required them to self-report to be cognitively healthy, which might have increased the observed longevity effect. Apart from the *APOE* variants, the AD-group included 15 variants, all of which were among the first to be associated with AD through GWAS (*CR1, CD33, BIN1, MS4A6A, PICALM,* and *SLC24A4*),[26, 27] eventually representing variants with the strongest effect on AD.Functional annotation showed significant enrichment of β-amyloid metabolism, which aligns with the importance of functional *APP* metabolism in maintaining brain health. We also observed functional enrichment of endocytosis and immune signaling, and a specific cell-type enrichment for microglia. This is in line with the currently growing hypothesis of the involvement of immune dysfunctions in the etiology underlying AD.[28, 29]

Variants with a larger effect on longevity than AD
The second-largest group of variants constituted a subset of 11 variants with a larger effect on longevity than the effect on AD, which suggests that these variants may be involved in other age-related diseases or general age-related processes. The AD-association of most of these variants is relatively recent, likely due to small effect sizes (ORs) or variants rareness (low minor allele frequency, MAF); both features necessitate a very large number of samples to identify these variants as significantly associating with the disease. The variants within this group were specifically enriched for immune response and endocytosis, which are known hallmarks of longevity.[1, 30, 31] In addition to the rare non-synonymous variant in the *PLCG2* gene (*rs72824905*, MAF:

0.6%), which was recently observed to be protective against AD, frontotemporal dementia (FTD) and dementia with Lewy bodies, other variants within this group were previously linked with disease risk factors. One of the two non-synonymous variants in the *SHARPIN* gene, variant *rs34173062* (MAF: 5.7%), has been associated with respiratory system diseases in GWAS.[32, 33, 34] Variant *rs7185636* (MAF: 17.1%), intronic of the IQCK gene, is in complete linkage with a variant (*rs7191155*, R^2=0.95), which was previously associated with body-mass index (BMI).[35] The variant *rs876461* (MAF: 13.0%) near the *PRKD3* gene is in linkage with variant *rs13420463* (R^2=0.42), which has been associated with systolic blood pressure.[36] Further, the variant near *CD2AP* gene associates with the development and maintenance of the blood-brain barrier, a specialized vascular structure of the central nervous system which, when disrupted, has been linked with epilepsy, stroke and AD.[37] Variant *rs9275152* (MAF: 10.4%) maps to the complex Human-Leukocyte-Antigen (HLA) region, which codes for cell-surface proteins responsible for the regulation of the adaptive immune system. In numerous GWAS, variants in the HLA region were associated with autoimmune diseases, cancer, and longevity.[6, 38] The AD-associated variant in this region (*rs9275152*) is also a risk variant for Parkinson's disease.[39] Finally, the genomic region surrounding the *SPI1* gene (in which variant *rs3740688* maps) has been previously associated with cognitive traits (intelligence, depression)[40] and, with lower evidence, with kidney disease and cancer.[41, 42] The remaining variants *rs56402156*, *rs7920721*, and *rs4351014* (in/near *EPHA1, ECHDC3*, and *HS3ST1*) have not been directly associated with other traits, although their associated genes were implicated in systemic lupus erythematosus (*HS3ST1*) and cancer (*EPHA1, ECHDC3*).[43, 44, 45] Together, these findings suggest that the counterpart of each risk-increasing allele, the AD-protective alleles, might give a survival advantage that is not only specific to AD. Their functional and cell-type annotations suggest that they contribute to the maintenance of regulatory stimuli in the immune and endosomal systems, which may be essential to maintain brain and overall physical health, necessary to reach extremely old ages in good cognitive health.[16]

Variants associated with increased risk of AD and increased longevity risk
Unexpectedly, ten variants increased the risk of AD while at the same time increasing the chance to reach ages over 100 in good cognitive health, which is an unexpected balance. We note that the *IED* distributions of these variants were broad, and in some cases even showed a bimodal behavior (in/near *KANSL1, IL34, CHRNE*): this is attributable to the small effect-sizes (and

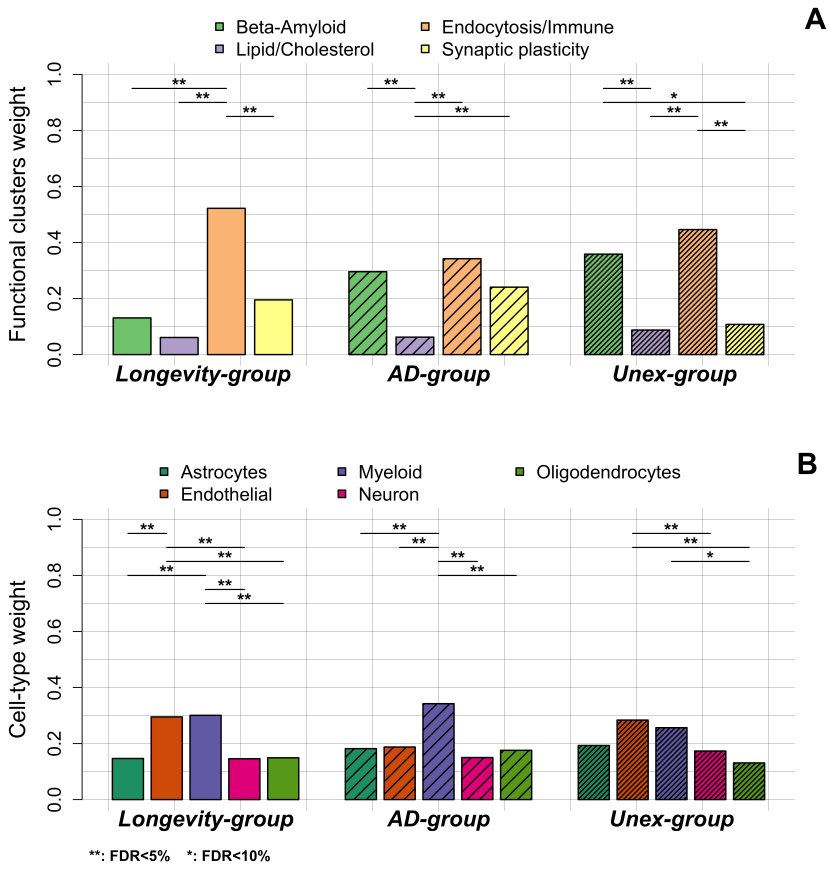


Figure 4.3: **Comparison of functional annotation and cell-type annotation within the Longevity-, AD- and Unex-groups. A.** The weights of the 4 functional clusters within the Longevity-, AD- and Unex-groups. **B.** The weights of the different cell-types in the brain, per group. Differences in functional weights and cell-type weights within each group were calculated using Wilcoxon sum rank tests. The resulting p-values were FDR-corrected.

large standard errors) on longevity for these variants, which caused data points to easily flip between the expected and unexpected direction during the sampling procedure. Replication of the direction of the variant effect in an independent dataset of parental longevity indicated that the unexpected direction was replicated in only the *CLU* and *CHRNE* variants, suggesting that future studies will have to further explore (the robustness of) these unexpected effects. One explanation for such counter-intuitive effects may be a variant interaction with other variants, which was shown for the variant in the *KANSL1* and *CLU* gene with respect to the *APOE* genotype.[46] Therefore, carrying the risk allele of such variants may specifically affect the risk of AD in *APOE* ϵ4 allele carriers, which are not prevalent among cognitively healthy centenarians. An alternative explanation may be that these variants have age-dependent effects: for example, high blood pressure at midlife increases the risk of AD, but after the age of 85 a high blood pressure protects against AD.[47] Similarly, a high body-mass-index (BMI) increases the risk of AD at midlife, while being protective at older ages.[48] In line with this hypothesis, the AD variant in/near *IL34* gene codes for a cytokine that is crucial for the differentiation and the maintenance of microglia.[49] Although further studies are needed, an excessive differentiation in middle-age individuals may increase brain-related inflammation and AD-risk, while it might compensate for the slower differentiation and immune activity at very old ages. Indeed, next to *IL34*, several genes that may be affected by these Unex-variants, such as *PTK2B* and *INPP5D*, play a role in aging-associated processes, such as cellular senescence or immunity.[50, 51]

4.4.4 Strengths and weaknesses

We acknowledge that our findings are based on relatively small sample sizes, especially for the cognitively healthy centenarian group. This phenotype is rare, and individuals need to be individually approached for study inclusion,[14] which is prohibitive for large sample collection. As population subjects in our comparison, we used individuals from five different cohorts: all from the same (Dutch) population, all tested cognitively intact, and did not convert to dementia at the time of analyses. It is known that the analysis of genetic variants with small effect-sizes in relatively small sample sizes leads to large confidence intervals: we took this uncertainty into account by bootstrapping effect sizes, causing the *IED* of several variants to be widely spread. By focusing our analysis on SNPs that were genome-wide significantly associated with AD (thus habving tight confidence intervals), we

limited this dispersion to the effects-sizes on longevity only. For this reason, we anticipate that using a random set of SNPs (*e.g.* to investigate the basic properties of the *IED*), would increase further more the dispersion of data points along the longevity-AD-unexpected spectrum, as confidence intervals on both longevity and AD would likely be larger. Although our work represents a first step towards understanding the effect of AD-associated variants on longevity, a replication analysis in larger cohorts of centenarians and/or long-lived individuals is warranted to further support our findings. Secondly, in the functional annotation analysis, we had to deal with the problem that the downstream effect of AD-associated variants is often unclear. To accommodate this uncertainty, we allower multiple genes to be associated with each variant. However, it is likely that our variant-pathways annotation will change as we gain more understanding about these variant-gene-effects, the likely affected genes, and their functions. When we inspected the parental-longevity GWAS, most of the variants that were in the expected direction in our study were also in the same direction in the GWAS; however, this was not true for all variants. The variant that deviated the most between our study and the parental-longevity GWAS was *rs9275152* in the HLA region: while we clustered this variant in the longevity-group, in the parental-longevity GWAS the direction of effect was opposite (*i.e.* unexpected), suggesting that the variant increased the risk of AD and at the same time the chance of a long lifespan. The genomic region to which HLA maps is biologically known to be affected by many recombination events and may be population- and environment-dependent, which may explain this divergence.[52] In addition to HLA-variant, variant *rs34674752* in the *SHARPIN* gene reported the second-largest effect-size in our study (after *APOE* ϵ4), while the effect-size of this variant in the GWAS was very small, yet in the expected direction. To this end, we note that the individuals used in the parental-longevity GWAS were themselves not extremely old individuals, such that possible pleiotropic effects at very old ages, as described earlier, may not be observable in this GWAS. However, while we observed overall consistency in effect-size direction for variants in the expected direction, 6/8 of the variants in the unexpected direction were in the expected direction in the GWAS, with variants near *SORL1, IL34*, and *ADAM10* having the most noticeable differences. We speculate that the relatively young ages of the GWAS samples, together with the small sample size of our centenarian cohort may be the cause of such discrepancy.

4.4.5 Conclusions

Most AD-associated variants that increase the risk of the disease are associated with lower odds of longevity. We identified a subset of variants with a larger effect on longevity than on AD, that were previously associated as risk-factors for other age-related diseases, and that are selectively enriched for endocytosis and immune signaling functions.

4.5 Acknowledgements

The following studies and consortia have contributed to this manuscript. Amsterdam Dementia Cohort (ADC): Research at the Alzheimer Center Amsterdam is part of the neurodegeneration research program of Amsterdam Neuroscience. 100-plus Study: we are grateful for the collaborative efforts of all participating centenarians and their family members and/or relatives. Wiesje van der Flier holds the Pasman chair. Longitudinal Aging Study of Amsterdam (LASA): the authors are grateful to all LASA participants, the fieldwork team, and all researchers for their ongoing commitment to the study. The Alzheimer Center Amsterdam is supported by Stichting Alzheimer Nederland and Stichting VUmc Fonds. The clinical database structure was developed with funding from Stichting Dioraphte. The SCIENCe project is supported by a research grant from Gieskes Strijbis Fonds. Genotyping of the Dutch case–control samples was performed in the context of EADB (European Alzheimer DNA Biobank), funded by the JPco-fuND FP-829-029 (ZonMW project number, 733051061). The 100-plus Study was supported by Stichting Alzheimer Nederland (WE09.2014-03), Stichting Diorapthe, horstingstuit foundation, Memorabel (ZonMW project number 733050814), and Stichting VUmc Fonds. Genotyping of the 100-plus Study was performed in the context of EADB (European Alzheimer DNA biobank) funded by the JPco-fuND FP-829-029 (ZonMW project number, 33051061). LASA is largely supported by a grant from the Netherlands Ministry of Health, Welfare and Sports, Directorate of Long-Term Care. **Conflict of interest:** all the authors in the study declared no conflict of interest. The funders had no role in the design of the study at any stage.

4.6 Full author list and affiliations

Niccolo' Tesi,[1,2,3] Marc Hulsman,[1,2,3] Sven J. van der Lee,[1,2] Iris E. Jansen,[1,4] Najada Stringa,[5] Natasja M. van Schoor,[5] Martijn Huisman,[5] Philip Scheltens,[1] Marcel J.T. Reinders,[3] Wiesje M. van der Flier,[1,5] and Henne Holstege [1,2,3]

[1] Alzheimer Centre, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[2] Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[3] Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

[4] Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, VU, Amsterdam, The Netherlands

[5] Department of Epidemiology and Data Sciences, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

4.7 Supplementary Methods

4.7.1 Populations

The 100-plus Study focuses on the biomolecular aspect of preserved cognitive health until extremely old ages. This study includes (1) Dutch-speaking centenarians who can (2) provide official evidence for being aged 100 years or older, (3) self-report to be cognitively healthy, which is confirmed by an informant (i.e. a child or close relation), (4) consent to donation of a blood sample and (5) consent to (at least) two home-visits from a researcher, which includes an interview and neuropsychological testing.[14] This study also includes (1) siblings or children from centenarians who participate in the 100-plus Study, or partners thereof who (2) agree to donate a blood sample, (3) agree to fill in a family history, lifestyle history, and disease history questionnaire. The Longitudinal Aging Study of Amsterdam (LASA) is an ongoing longitudinal study of older adults initiated in 1991, with the main objective to determine predictors and consequences of aging.[18, 19] The SCIENCe is a prospective cohort study of subjective cognitive decline (SCD) patients.[20, 53] Participants undergo extensive assessment, including cerebrospinal fluid collection (CSF) and optional amyloid positron emission tomography scan (PET), with annual follow-up. The primary outcome measure is clinical progression. All individuals were labeled cognitively intact. The Netherlands Brain Bank (NBB) cohort is a prospective donor program for psychiatric diseases. All subjects were labeled cognitively intact after neuropathological examination.[22] The Netherland Twin Registry study (NTR) was established in 2004 to collect biological and environmental data in twin families to create a resource for genetic studies on health, lifestyle, and personality.[23]

4.7.2 Genotyping and imputation

Genetic variants in our populations were determined by standard genotyping and imputation methods, and we applied established quality control methods: we genotyped all individuals with the Illumina Global Screening Array (GSAsharedCUSTOM-20018389-A2) and excluded individuals with low-quality genotypes (individual call rate <98%, variant call rate <98%), individuals with sex mismatches and variants deviating from Hardy-Weinberg equilibrium ($p < 1x10^{-6}$). Genotypes were prepared for imputation comparing variants identifiers, strand and allele frequencies to the Haplotype Reference Panel (HRC v1.1, April 2016), and all remaining variants were submitted to the Sanger imputation server (`https://imputation.sanger.ac.uk`).[54] The server uses EAGLE2 (v2.0.5) to phase the data, and imputation to the ref-

erence panel was performed with PBWT.[55, 56] Before analysis, we excluded individuals of non-European ancestry and individuals with a family relation, leaving 2,905 population subjects and 343 cognitively healthy centenarians for the analysis.

4.7.3 Variant annotation

Variant-gene mapping
We annotated each variant to the likely affected gene(s), so-called *variant-gene mapping*, combining annotation from Combined Annotation Dependent Depletion (CADD, v1.3), expression-quantitative-trait-loci in the blood (eQTL from GTEx v8), and positional mapping (from RefSeq build 98).[57, 58, 59] In the case of coding variants, we confidently associated the variant with the corresponding gene. Alternatively, we first considered possible eQTL associations. When these were not available, we included all genes at increasing distance d from the variant (starting with $d \leq 50kb$, up to $d \leq 500kb$, increasing by $50kb$ until at least 1 gene was found). Our procedure allows the association of each variant with one or multiple genes (Figure 4.5).

Gene-pathway mapping
The resulting list of genes was used to find the molecular pathways enriched in the AD variants. See Figure 4.5 for a schematic representation of our annotation framework. We realized that allowing multiple genes to associate with each variant could result in an enrichment bias, as neighboring genes are often functionally related. To control this, we implemented a sampling technique: at each iteration, we (*i*) sampled one gene from the pool of genes associated with each variant, and (*ii*) performed a gene-set enrichment analysis with the resulting list of genes. The gene-set enrichment analysis was performed considering biological processes (BP) and implemented with the *enrichGO* function of the R package *clusterProfiler*, with all genes as background and correcting *p*-values controlling the False Discovery Rate (FDR). Finally, we averaged p-values for each enriched term over the iterations (N=1,000). To facilitate interpretation, we merged significantly enriched biological processes. First, we calculated the semantic similarity between all significant biological processes (*i.e.* FDR<5%) using Lin as a distance measure.[60] We then applied hierarchical clustering on the resulting distance matrix and selected the number of functional clusters using the dynamic tree-cut method as implemented in *cutreeDynamic* function from the R package *WGCNA*, specifying 15 as the minimum number of terms per cluster (using the default value of 20 resulted in 2 functional clusters only). To provide

an interpretation of each functional cluster, we selected the most frequent words describing the biological processes underlying each cluster, and show this as word-clouds as implemented in R package *wordcloud2*. Finally, by counting how often a functional cluster was associated with a gene, we could calculate a weighted annotation of each gene to the 4 functional clusters, so-called gene-pathway mapping (Figure 4.5). The variant-gene mapping as well as the gene-pathway mapping procedures were performed using the web-server application available at `https://snpxplorer.net`. [61] Due to the initial selection of significantly enriched BP, not every gene in the list of variant-associated genes is annotated with (at least one of) these terms. Consequently, these genes could not be related to the final functional clusters. To overcome this, we connect these genes to the functional clusters using a *k*-nearest neighbor (*k-NN*) imputation. The *k-NN* model was initially trained using the functional clusters as classes and the semantic similarity matrix between the enriched biological processes as features (feature terms). Then, for each gene with missing annotation, we (*i*) extracted all the biological processes the gene is involved in (input biological processes), and (*ii*) calculated the semantic similarity matrix between these terms and the feature terms, which defines the similarity between the input biological processes and the feature terms. Finally, we (*iii*) predicted the probability of classification of the similarity matrix to the classes (functional clusters), and used this as weight for the gene-pathway mapping (Figure 4.5).

Variant-pathway mapping

The variant-pathway mapping represents the combined annotation of each variant to the different functional clusters. As such, it depends on the variant-gene mapping and the gene-pathways mapping. Briefly, given a variant k, we (*i*) retrieved all the genes that were associated with the variant in the variant-gene mapping, G_k, and (*ii*) retrieved all the biological processes (gene ontology term identifiers) that were associated with these genes, GO_G. Because we clustered biological processes into functional clusters, by looking at which functional clusters the GO_G belonged to, we could assign a weight of association for variant k to each of the functional clusters.

4.7.4 Variant-cell-type mapping

To study brain-specific cell-types and their relationship with AD-associated variants, we used the publicly available gene expression dataset *GSE73721*: this dataset includes gene expression values of 6 fetal astrocyte samples, 12 adult astrocyte samples, 8 sclerotic hippocampal samples, 4 whole human

cortex samples, 4 adult mouse astrocyte samples, and 11 human samples of other purified central-nervous-system (CNS) cell-types. We restricted to the gene expression of 12 astrocyte samples and 11 samples of purified CNS cell-types from the cortex of adult humans (total N=23, mean age of 41.5±19.6 years). To calculate the variant-cell-type mapping, we averaged the gene expression of the genes mapping to the same variant.

4.8 Supplementary Figures

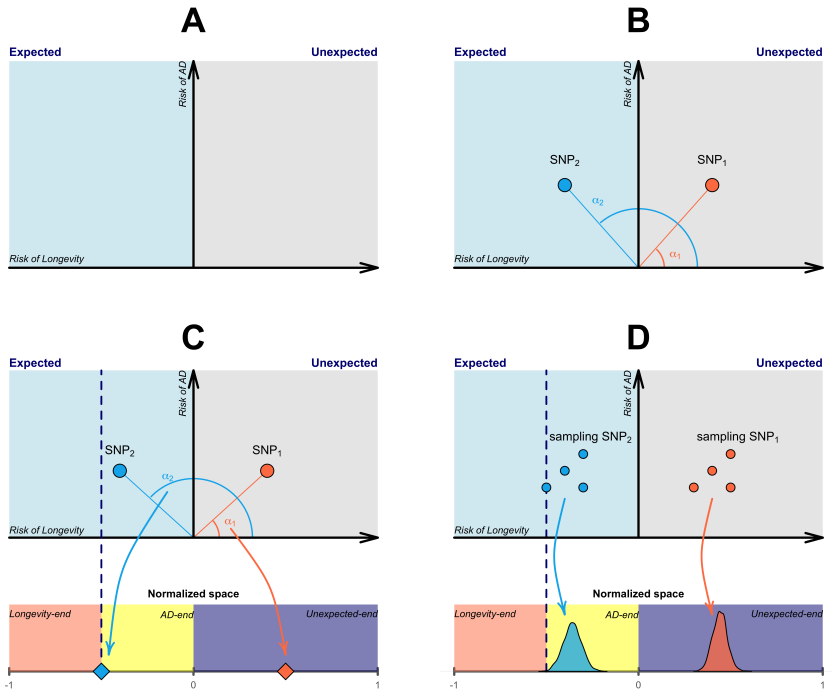


Figure 4.4: **Explanation of the distribution of imbalance variant effect direction (IED).** The figure shows the sequential steps for constructing the distribution of the expected direction of variant effect for AD-risk compared to longevity for two toy variants (SNP_1 and SNP_2). **A.** Axes definition, with the y-axis being the effect-size for AD-risk (log of odds ratio) of a variant, derived from literature and set positive by definition. The x-axis identifies the effect-size of a variant on longevity. This can be either positive or negative depending on the variant's association in cognitively healthy centenarians as opposed to population subjects. The blue area represents that the two effects are in the expected direction with respect to each other, *i.e.* a variant increases the risk of AD and at the same time decreases the chance of longevity. Oppositely, the grey area refers to the unexpected direction of effect. **B.** Two toy variants (SNP_1 and SNP_2) are shown as data points. $\alpha_{(1-2)}$ represents the angle of the data point vector with the x-axis. **C.** Normalization of the $\alpha_{(1-2)}$ value into an arbitrary space. Here, we used $[-1;1]$. **D.** Repeating this procedure for each bootstrap iteration of each variant, we obtained the distribution of imbalance effect direction for each variant (*IED*). Values smaller than 0 indicate the expected direction of effect, whereas values larger than 0 refer to the unexpected direction of effects. Additionally, values close to 0 indicate a larger AD effect than longevity effect, and values close to -1 suggest that the variant's longevity effect is larger than the AD effect.

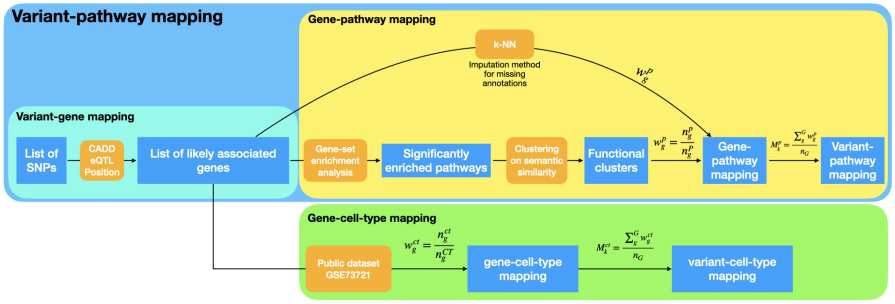


Figure 4.5: **Schematic representation of the variant-pathway and variant-cell-type mapping.** The figure shows a schematic representation of the annotation framework used to functionally annotate AD-associated variants and perform cell-type enrichment. Outputs are represented as blue squares, while methods are represented in orange. In the variant-gene mapping, showed in the grey box, we start from a list of variants and, through the integration of predicted variant consequences (CADD), eQTL and position, we obtain a list of genes. Note that here multiple genes may be associated with each variant. The yellow box shows the gene-pathway mapping: briefly, we perform gene-set enrichment analysis followed by clustering of the significantly enriched pathways to obtain functional clusters. We then calculate the gene-pathway mapping by looking at the (enriched) pathways associated with each gene and their associated functional clusters to get a weight for each gene-functional cluster association. Finally, we average the gene-pathway mapping of each gene associated with the same variant. Imputation methods (*k-NN*) are implemented for genes with missing annotation to obtain the gene-pathway mapping. Together, the grey box and the yellow box form the variant-pathway mapping. At the bottom, the green box shows the gene-cell-type enrichment using the public dataset *GSE73721* of gene expression in different brain cell-types. Similar to the gene-pathway mapping, we calculate a weight of association of each gene to each cell-type, and we average these weights in case multiple genes mapped to the same variant (*variant-cell-type mapping*).

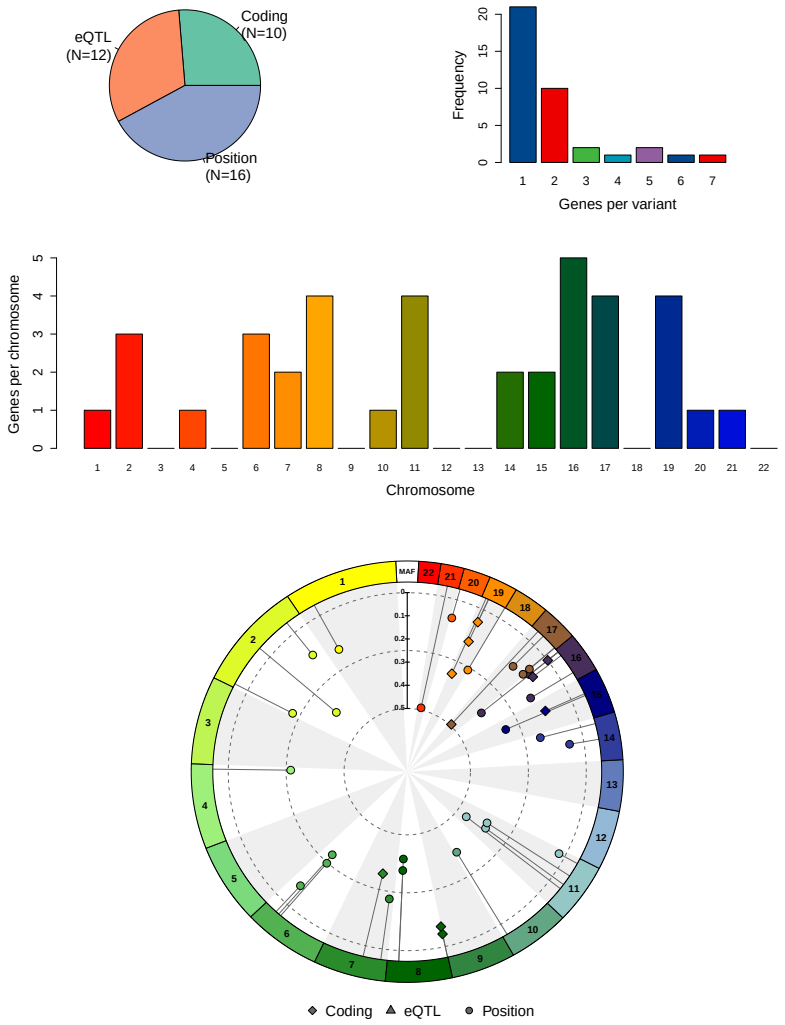


Figure 4.6: **Variant-gene mapping for the 38 AD-associated variants. A.** The sources used to annotate each variant to the likely affected genes. *Coding*: variants located in the coding region of a gene (*e.g.* synonymous or non-synonymous variants). *eQTL*: variants associated with RNA expression changes in blood from the GTEx consortium. *Position*: variants intronic or intergenic without evidence of eQTL associations that were annotated based on neighboring genes. **B.** Barplot of the number of genes associated with each variant. **C.** Distribution of genes across the chromosomes. **D.** Distribution of the previously identified variants along the genome together with each variant's minor allele frequency and annotation.

4.9 Supplementary Material

Supplementary Tables can be accessed by scanning the following code or accessing the journal's website here.

References

[1] Linda Partridge, Joris Deelen, and P. Eline Slagboom. "Facing up to the global challenges of ageing". In: *Nature* 561.7721 (Sept. 2018), pp. 45–56. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0457-8.

[2] David Melzer, Luke C. Pilling, and Luigi Ferrucci. "The genetics of human ageing". In: *Nature Reviews Genetics* (Nov. 2019). ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-019-0183-6.

[3] Thomas T. Perls et al. "Life-long sustained mortality advantage of siblings of centenarians". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.12 (June 2002), pp. 8442–8447. ISSN: 0027-8424. DOI: 10.1073/pnas.122587599.

[4] Graziella Caselli et al. "Family clustering in Sardinian longevity: a genealogical approach". In: *Experimental Gerontology* 41.8 (Aug. 2006), pp. 727–736. ISSN: 0531-5565. DOI: 10.1016/j.exger.2006.05.009.

[5] Joris Deelen et al. "A meta-analysis of genome-wide association studies identifies multiple longevity genes". In: *Nature Communications* 10.1 (Dec. 2019). ISSN: 2041-1723. DOI: 10.1038/s41467-019-11558-2.

[6] Paul RHJ Timmers et al. "Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances". In: *eLife* 8 (Jan. 2019). ISSN: 2050-084X. DOI: 10.7554/eLife.39856.

[7] "2012 Alzheimer's disease facts and figures". In: *Alzheimer's & Dementia* 8.2 (Mar. 2012), pp. 131–168. ISSN: 15525260. DOI: 10.1016/j.jalz.2012.02.001.

[8] María M. Corrada et al. "Dementia incidence continues to increase with age in the oldest old: The 90+ study". In: *Annals of Neurology* 67.1 (Jan. 2010), pp. 114–121. ISSN: 03645134, 15318249. DOI: 10.1002/ana.21915.

[9] Margaret Gatz et al. "Role of genes and environments for explaining Alzheimer disease". In: *Archives of General Psychiatry* 63.2 (Feb. 2006), pp. 168–174. ISSN: 0003-990X. DOI: 10.1001/archpsyc.63.2.168.

[10] Alzheimer Disease Genetics Consortium (ADGC), et al. "Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing". In: *Nature Genetics* 51.3 (Mar. 2019), pp. 414–430. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-019-0358-2.

[11] Iris E. Jansen et al. "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk". In: *Nature Genetics* 51.3 (Mar. 2019), pp. 404–413. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-018-0311-9.

[12] Rebecca Sims et al. "Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease". In: *Nature Genetics* 49.9 (Sept. 2017), pp. 1373–1384. ISSN: 1546-1718. DOI: 10.1038/ng.3916.

[13] Itziar de Rojas et al. *Common variants in Alzheimer's disease: Novel association of six genetic variants with AD and risk stratification by polygenic risk scores*. preprint. Genetic and Genomic Medicine, Nov. 2019. DOI: 10.1101/19012021.

[14] Henne Holstege et al. "The 100-plus Study of Dutch cognitively healthy centenarians: rationale, design and

cohort description". In: (Apr. 2018). DOI: 10.1101/295287.

[15] Niccolò Tesi et al. "Centenarian controls increase variant effect sizes by an average twofold in an extreme case–extreme control analysis of Alzheimer's disease". In: *European Journal of Human Genetics* (Sept. 2018). ISSN: 1018-4813, 1476-5438. DOI: 10.1038/s41431-018-0273-5.

[16] Niccolò Tesi et al. "Immune response and endocytosis pathways are associated with the resilience against Alzheimer's disease". In: *Translational Psychiatry* 10.1 (Dec. 2020), p. 332. ISSN: 2158-3188. DOI: 10.1038/s41398-020-01018-7.

[17] Natalya Ponomareva et al. "Age-dependent effect of Alzheimer's risk variant of CLU on EEG alpha rhythm in non-demented adults". In: *Frontiers in Aging Neuroscience* 5 (2013). ISSN: 1663-4365. DOI: 10.3389/fnagi.2013.00086.

[18] Emiel O. Hoogendijk et al. "The Longitudinal Aging Study Amsterdam: cohort update 2016 and major findings". In: *European Journal of Epidemiology* 31.9 (Sept. 2016), pp. 927–945. ISSN: 0393-2990, 1573-7284. DOI: 10.1007/s10654-016-0192-0.

[19] M. Huisman et al. "Cohort Profile: The Longitudinal Aging Study Amsterdam". In: *International Journal of Epidemiology* 40.4 (Aug. 2011), pp. 868–876. ISSN: 0300-5771, 1464-3685. DOI: 10.1093/ije/dyq219.

[20] Rosalinde E. R. Slot et al. "Subjective Cognitive Impairment Cohort (SCIENCe): study design and first results". In: *Alzheimer's Research & Therapy* 10.1 (Dec. 2018). ISSN: 1758-9193. DOI: 10.1186/s13195-018-0390-y.

[21] Wiesje M. van der Flier and Philip Scheltens. "Amsterdam Dementia Cohort: Performing Research to Optimize Care". In: *Journal of Alzheimer's Disease* 62.3 (Mar. 2018). Ed. by George Perry, Jesus Avila, and Xiongwei Zhu, pp. 1091–1111. ISSN: 13872877, 18758908. DOI: 10.3233/JAD-170850.

[22] Marleen C. Rademaker, Geertje M. de Lange, and Saskia J.M.C. Palmen. "The Netherlands Brain Bank for Psychiatry". In: *Handbook of Clinical Neurology*. Vol. 150. Elsevier, 2018, pp. 3–16. ISBN: 978-0-444-63639-3. DOI: 10.1016/B978-0-444-63639-3.00001-3.

[23] Gonneke Willemsen et al. "The Netherlands Twin Register Biobank: A Resource for Genetic Epidemiological Studies". In: *Twin Research and Human Genetics* 13.3 (June 2010), pp. 231–245. ISSN: 1832-4274, 1839-2628. DOI: 10.1375/twin.13.3.231.

[24] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *American Journal of Human Genetics* 81.3 (Sept. 2007), pp. 559–575. ISSN: 0002-9297. DOI: 10.1086/519795.

[25] Hui Shi et al. "Genetic variants influencing human aging from late-onset Alzheimer's disease (LOAD) genome-wide association studies (GWAS)". In: *Neurobiology of Aging* 33.8 (Aug. 2012), 1849.e5–1849.e18. ISSN: 01974580. DOI: 10.1016/j.neurobiolaging.2012.02.014.

[26] Denise Harold et al. "Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease". In: *Nature Genetics* 41.10 (Oct. 2009), pp. 1088–1093. ISSN: 1546-1718. DOI: 10.1038/ng.440.

[27] Jean-Charles Lambert et al. "Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease". In: *Nature Genetics* 41.10 (Oct. 2009), pp. 1094–1099. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.439.

[28] Anastasia G. Efthymiou and Alison M. Goate. "Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk". In: *Molecular Neurodegeneration* 12.1 (Dec. 2017). ISSN: 1750-1326. DOI: 10.1186/s13024-017-0184-x.

[29] David V. Hansen, Jesse E. Hanson, and Morgan Sheng. "Microglia in Alzheimer's disease". In: *The Journal of Cell Biology* 217.2 (Feb. 2018), pp. 459–472. ISSN: 0021-9525, 1540-8140. DOI: 10.1083/jcb.201709069.

[30] Amir A. Sadighi Akha. "Aging and the immune system: An overview". In: *Journal of Immunological Methods* 463 (Dec. 2018), pp. 21–26. ISSN: 00221759. DOI: 10.1016/j.jim.2018.08.005.

[31] Santiago Solé-Domènech et al. "The endocytic pathway in microglia during health, aging and Alzheimer's disease". In: *Ageing Research Reviews* 32 (Dec. 2016), pp. 89–103. ISSN: 15681637. DOI: 10.1016/j.arr.2016.07.002.

[32] William J. Astle et al. "The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease". In: *Cell* 167.5 (2016), 1415–1429.e19. ISSN: 1097-4172. DOI: 10.1016/j.cell.2016.10.042.

[33] Thorunn A. Olafsdottir et al. "Eighty-eight variants highlight the role of T cell regulation and airway remodeling in asthma pathogenesis". In: *Nature Communications* 11.1 (2020), p. 393. ISSN: 2041-1723. DOI: 10.1038/s41467-019-14144-8.

[34] Gleb Kichaev et al. "Leveraging Polygenic Functional Enrichment to Improve GWAS Power". In: *American Journal of Human Genetics* 104.1 (2019), pp. 65–75. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2018.11.008.

[35] Thomas J. Hoffmann et al. "A Large Multiethnic Genome-Wide Association Study of Adult Body Mass Index Identifies Novel Loci". In: *Genetics* 210.2 (2018), pp. 499–515. ISSN: 1943-2631. DOI: 10.1534/genetics.118.301479.

[36] Helen R. Warren et al. "Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk". In: *Nature Genetics* 49.3 (Mar. 2017), pp. 403–415. ISSN: 1546-1718. DOI: 10.1038/ng.3768.

[37] J. Nicholas Cochran et al. "The Alzheimer's disease risk factor CD2AP maintains blood–brain barrier integrity". In: *Human Molecular Genetics* 24.23 (Dec. 1, 2015), pp. 6667–6674. ISSN: 0964-6906, 1460-2083. DOI: 10.1093/hmg/ddv371.

[38] Gergely Bodis, Victoria Toth, and Andreas Schwarting. "Role of Human Leukocyte Antigens (HLA) in Autoimmune Diseases". In: *Rheumatology and Therapy* 5.1 (June 2018), pp. 5–20. ISSN: 2198-6576, 2198-6584. DOI: 10.1007/s40744-018-0100-z.

[39] Sara Bandres-Ciga et al. "The Genetic Architecture of Parkinson Disease in Spain: Characterizing Population-Specific Risk, Differential Haplotype Structures, and Providing Etiologic Insight". In: *Movement Disorders: Official Journal of the Movement Disorder*

Society 34.12 (2019), pp. 1851–1863. ISSN: 1531-8257. DOI: 10.1002/mds.27864.

[40] Gail Davies et al. "Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function". In: *Nature Communications* 9.1 (Dec. 2018). ISSN: 2041-1723. DOI: 10.1038/s41467-018-04362-x.

[41] Cristian Pattaro et al. "Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function". In: *Nature Communications* 7 (Jan. 21, 2016), p. 10023. ISSN: 2041-1723. DOI: 10.1038/ncomms10023.

[42] Kyriaki Michailidou et al. "Association analysis identifies 65 new breast cancer risk loci". In: *Nature* 551.7678 (2017), pp. 92–94. ISSN: 1476-4687. DOI: 10.1038/nature24284.

[43] Carl D. Langefeld et al. "Transancestral mapping and genetic load in systemic lupus erythematosus". In: *Nature Communications* 8 (2017), p. 16021. ISSN: 2041-1723. DOI: 10.1038/ncomms16021.

[44] N I Herath et al. "Epigenetic silencing of EphA1 expression in colorectal cancer is correlated with poor survival". In: *British Journal of Cancer* 100.7 (Apr. 2009), pp. 1095–1102. ISSN: 0007-0920, 1532-1827. DOI: 10.1038/sj.bjc.6604970.

[45] Sajjad Rafiq et al. "A genome wide meta-analysis study for identification of common variation associated with breast cancer prognosis". In: *PloS One* 9.12 (2014), e101488. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0101488.

[46] G. Jun et al. "A novel Alzheimer disease locus located near the gene encoding tau protein". In: *Molecular Psychiatry* 21.1 (Jan. 2016), pp. 108–117. ISSN: 1476-5578. DOI: 10.1038/mp.2015.23.

[47] Emer R. McGrath et al. "Blood pressure from mid- to late life and risk of incident dementia". In: *Neurology* 89.24 (Dec. 12, 2017), pp. 2447–2454. ISSN: 1526-632X. DOI: 10.1212/WNL.0000000000004741.

[48] W. L. Xu et al. "Midlife overweight and obesity increase late-life dementia risk: a population-based twin study". In: *Neurology* 76.18 (May 3, 2011), pp. 1568–1574. ISSN: 1526-632X. DOI: 10.1212/WNL.0b013e3182190d09.

[49] Yaming Wang and Marco Colonna. "Interkeukin-34, a cytokine crucial for the differentiation and maintenance of tissue resident macrophages and Langerhans cells: Highlights". In: *European Journal of Immunology* 44.6 (June 2014), pp. 1575–1581. ISSN: 00142980. DOI: 10.1002/eji.201344365.

[50] S. J. Ryu et al. "Role of Src-specific phosphorylation site on focal adhesion kinase for senescence-associated apoptosis resistance". In: *Apoptosis* 11.3 (Mar. 2006), pp. 303–313. ISSN: 1360-8185, 1573-675X. DOI: 10.1007/s10495-006-3978-9.

[51] Samantha D. Pauls and Aaron J. Marshall. "Regulation of immune cell signaling by SHIP1: A phosphatase, scaffold protein, and potential therapeutic target". In: *European Journal of Immunology* 47.6 (June 2017), pp. 932–945. ISSN: 00142980. DOI: 10.1002/eji.201646795.

[52] Timothy A. Jinam. "Human Leukocyte Antigen (HLA) Region in Human Population Studies". In: *Evolution of the Human Genome I*. Ed. by Naruya Saitou. Series Title: Evolutionary Studies. Tokyo: Springer Japan, 2017, pp. 173–179. ISBN: 978-4-431-56601-4 978-4-431-56603-8. DOI: 10.1007/978-4-431-56603-8_9.

[53] Wiesje M. van der Flier et al. "Optimizing patient care and research: the Amsterdam Dementia Cohort". In: *Journal of Alzheimer's disease: JAD* 41.1 (2014), pp. 313–327. ISSN: 1875-8908. DOI: 10.3233/JAD-132306.

[54] Shane McCarthy et al. "A reference panel of 64,976 haplotypes for genotype imputation". In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1279–1283. ISSN: 1546-1718. DOI: 10.1038/ng.3643.

[55] Richard Durbin. "Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT)". In: *Bioinformatics (Oxford, England)* 30.9 (May 1, 2014), pp. 1266–1272. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btu014.

[56] Po-Ru Loh et al. "Reference-based phasing using the Haplotype Reference Consortium panel". In: *Nature Genetics* 48.11 (2016), pp. 1443–1448. ISSN: 1546-1718. DOI: 10.1038/ng.3679.

[57] Philipp Rentzsch et al. "CADD: predicting the deleteriousness of variants throughout the human genome". In: *Nucleic Acids Research* 47 (D1 Jan. 2019), pp. D886–D894. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gky1016.

[58] Nuala A. O'Leary et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic Acids Research* 44 (D1 Jan. 2016), pp. D733–745. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1189.

[59] GTEx Consortium. "The Genotype-Tissue Expression (GTEx) project". In: *Nature Genetics* 45.6 (June 2013), pp. 580–585. ISSN: 1546-1718. DOI: 10.1038/ng.2653.

[60] Bridget T. McInnes and Ted Pedersen. "Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text". In: *Journal of Biomedical Informatics* 46.6 (Dec. 2013), pp. 1116–1124. ISSN: 15320464. DOI: 10.1016/j.jbi.2013.08.008.

[61] Niccolo Tesi et al. "snpXplorer: a web application to explore human SNP-associations and annotate SNP-sets". In: *Nucleic Acids Research* 49.W1 (May 2021), W603–W612. ISSN: 0305-1048. DOI: 10.1093/nar/gkab410.

5. Genetic predisposition to longevity

Polygenic risk score of longevity predicts longer survival across an age-continuum

Niccolo' Tesi, Sven J. van der Lee, Marc Hulsman, Iris E. Jansen, Najada Stringa, Natasja M. van Schoor, Martijn Huisman, Philip Scheltens, Wiesje M. van der Flier, Marcel J.T. Reinders, and Henne Holstege

This chapter was published in *The Journal of Gerontology: Series A*
https://doi.org/10.1093/gerona/glaa289

Abstract

Studying the genome of centenarians may give insights into the molecular mechanisms underlying extreme human longevity and the escape of age-related diseases. Here, we set out to construct polygenic-risk-scores (PRS) for longevity and to investigate the functions of longevity-associated variants. Using a cohort of centenarians with maintained cognitive health (N=343), a population-matched cohort of older-adults from five cohorts (N=2,905), and summary statistics data from a GWAS on parental longevity, we constructed a PRS including 330 variants that significantly discriminated between centenarians and older-adults. This PRS was also associated with longer survival in an independent sample of younger individuals, (p=0.02), leading up to a 4-year difference in survival based on common genetic factors only. We show that this PRS was, in part, able to compensate for the deleterious effect of the *APOE-*ϵ4 allele. Using an integrative framework, we annotated the 330 variants included in this PRS by the genes they associate with. We find that they are enriched with genes associated with cellular differentiation, developmental processes, and cellular response to stress. Together, our results indicate that an extended human lifespan is, in part, the result of a constellation of variants each exerting small advantageous effects on aging-related biological mechanisms that maintain overall health and decrease the risk of age-related diseases.

5.1 Introduction

The human aging process is influenced by genetic and environmental factors, which makes it one of the most complex traits to study.[1, 2] Previous studies estimated that the heritability of lifespan up to ∼70 years of age ranges 10-25%.[3, 4] However, to reach higher ages we become increasingly dependent on the favorable genetic elements of our genomes. In fact, the heritability of becoming a centenarian has been estimated to be 60%.[5] Interestingly, centenarian genomes are depleted of single-nucleotide-polymorphisms (SNPs) associated with age-related diseases, while they are enriched with protective SNPs.[6, 7] Therefore, studying the genetic variants enriched in centenarians may give insights into the underlying etiology of extreme human longevity.[6, 7]

The research of SNPs that influence the human lifespan has focused mainly on the replication of candidate genes discovered in model organisms.[8, 9] Recently, genome-wide association studies (GWAS) have been performed to identify genetic loci associated with longevity. GWAS of longevity, in which the frequency of genetic variants is compared between long-lived persons and the average population, do not require prior knowledge and have the potential to discover new genetic determinants.[10] These studies have identified a constellation of SNPs associated with a longer lifespan across a wide range of populations.[11, 12, 13, 14, 15, 16] However, the association of the identified genetic loci has typically a low replication rate across independent studies, with only the *APOE-*ϵ4 allele (variant *rs429358*) and genetic variants in *CDKN2A/B* gene consistently associated with reduced lifespan.[11, 14, 15, 17] The difficulty in replicating longevity-associated SNPs may be attributable to different measures of survival and longevity, different statistical methods, and population dynamics.[8, 15, 18] For example, some studies used a dichotomous longevity phenotype based on the survival to ages above 90 or 100 years, others used the top 10% or 1% of survivors in a population,[12, 14] while other studies modeled age at death as a continuous variable and yet others used more sophisticated statistical models.[13, 15] On top of methodological and phenotypical divergencies between studies, population dynamics including gene-environmental interactions and population biases may potentially have a large effect on longevity,[18] and might explain the poor replication rate in independent cohorts. Lastly, the genetic variants identified thus far carry small effects, such that large sample sizes are required for an association with longevity to reach statistical significance in a GWAS setting.[8] Although poorly replicated, 29 genomic regions have

been associated with a longer lifespan in the most recent GWAS studies.[11, 12, 14, 15, 16, 18] The genes that harbor these variants have been implicated in age-related diseases including cardiovascular diseases (*APOE*, *ANRIL*), type I diabetes (*FOXO3*, *LPA*), cancer (*CDKN2B*, *BEND4*), and neurological diseases (*APOE*, *GPR78*, *GRIK2*).[13, 15] Together, this suggests that an extended human lifespan is associated with a lower genetic risk of age-related diseases.[8, 15, 19] Indeed, centenarians across populations have been shown to compress their disability period to the very end of their lives, escaping or delaying age-related diseases until extreme ages.[5, 20, 21, 22]

We hypothesize that variants associated with longevity are maximally enriched in cognitively healthy centenarians because, in addition to reaching at least 100 years (1% of the population), these centenarians are cognitively healthy, and represent an even smaller percentage of the general population (∼0.1%).[20] We previously found that the selection for cognitive health next to being 100 years or older is associated with prolonged longevity in this cohort compared to centenarians from the general population.[20, 21, 22, 23] Therefore, the centenarians in this cohort represent the ideal group to construct and test polygenic risk scores for longevity. A polygenic risk score (PRS) is a weighted score of independent variants representative of the risk to develop a phenotypic trait and can be used to study the combined influence of genetic factors on a certain trait. Although a PRS of parental longevity was previously associated with survival, validation in a cohort of extremely old individuals is missing. Besides, to prioritize SNPs to include in the PRS using a cohort of cognitively healthy agers may improve association statistics of the PRS. In this study, we started from 29 genomic regions previously associated with longevity: we annotated SNPs to likely affected genes and sought to detect significant associations using gene-based tests as opposed to single variant associations. Importantly, we constructed polygenic risk scores (PRS) combining the effect of multiple variants and tested the association of the risk scores (*i*) with becoming a cognitively healthy centenarian, and (*ii*) with survival in a subset of controls with follow-up data. We further explore the relationship between the PRS and the deleterious effect of *APOE-ϵ*4 allele, and using an innovative framework, we functionally annotate the variants included in the best PRS model.

5.2 Results

5.2.1 Study population

We studied the genetics underlying extreme human longevity in a case-control setting using as cases individuals that reached at least 100 years of age and who self-reported as cognitively healthy. As controls, we used a sample of population-matched, older-adults drawn from five different studies (see section 5.4). After establishing quality control of the genotyping data, 343 cognitively healthy centenarians (mean age at inclusion 101.4±1.8, 71.7% females) and 2,905 controls (mean age 68.3±11.5, 48.2 % females) were included in the analyses (Table S1).

5.2.2 Linking genetic variants with genes

We linked genetic variants previously associated with longevity (Table S2) to their likely affected genes. However, for non-coding variants, the closest gene is not necessarily the affected gene. Of the 29 investigated variants, only a few are coding (*N*=5), while most are intronic (*N*=16) or intergenic (*N*=8), for which variant consequences are unclear. To investigate the variant-effect on gene function, we combined variant consequences as predicted by the *Combined Annotation Dependent Depletion* (CADD),[24] *expression quantitative-trait-loci* (eQTL) in blood from the Genotype-Tissue Expression (GTEx) consortium,[25] and positional information to associate each variant to the gene(s) it likely affects. This allows each genetic variant to associate with one or more genes, depending on annotation certainties. With this procedure, the 29 genetic variants mapped to 65 unique genes: 16 SNPs mapped to 1 gene, while 6 mapped to 2 genes, 4 to 3 genes, 1 to 6 genes, 1 to 8 genes, and 1 to 12 genes (Figure 5.5 and Table S3). This annotation tool is freely accessible to the community at `https://snpxplorer.eu.ngrok.io`.

5.2.3 Combined association of multiple variants at the gene level

While single variant associations represent the standard procedure for GWAS, we hypothesized that testing the aggregated association of multiple variants across a gene might improve association statistics. In total, we tested the joint-association of variants at the gene-level for 53/65 genes using the MAGMA statistical framework (see section 5.4).[26] After correction for multiple tests (*False Discovery Rate*, FDR), the association of *APOE* and *CDKN2B* genes remained significant at FDR<10% (p=3.14x10^{-12} and p=0.002, respectively, Figure 5.6 and Table S4).

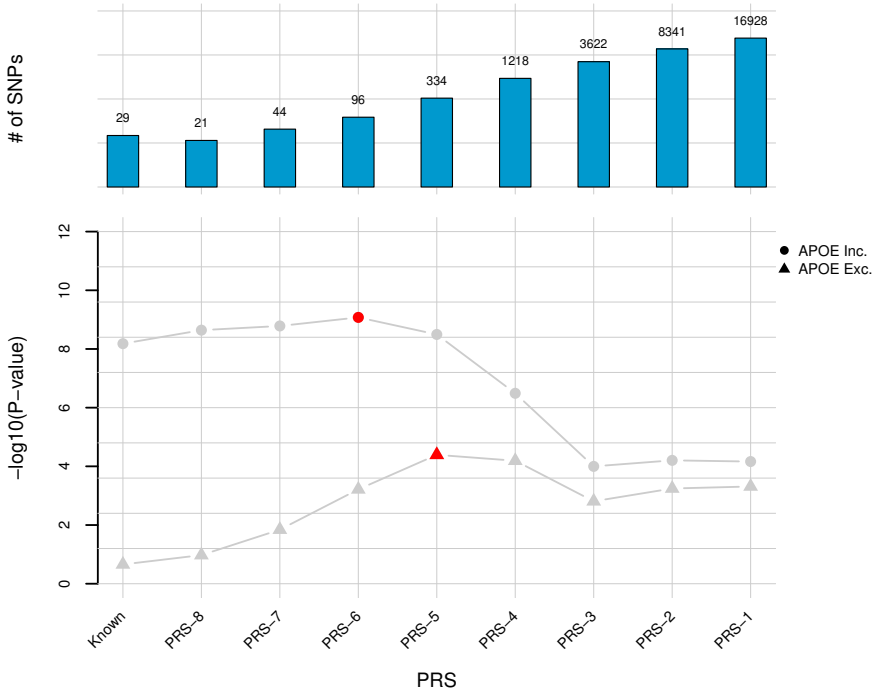


Figure 5.1: **Association of the PRSs with becoming a cognitively healthy centenarian and the effect of inclusion of sub-significant variants.** The top panel shows the number of variants (in $\log_{10}$ scale) included in each PRS (including *APOE* variants). The bottom panel shows the p-value of the difference in PRS score between the cognitively healthy centenarians and controls. Circles denote PRSs including the *APOE* variants and triangles PRSs without the *APOE* variants. Red points refer to the most significant models including and excluding *APOE*, respectively. Known refers to the PRS including the 29 previously identified variants associated with longevity. *PRS-x* refers to the PRS including additional sub-significant variants depending on the *p*-value of association of the variants in the by-proxy GWAS on longevity: PRS-8 for variants with $p<5\text{x}10^{-8}$; PRS-7 for variants with $p<5\text{x}10^{-7}$; PRS-6 for variants with $p<5\text{x}10^{-6}$, *etc.*

5.2.4 Polygenic Risk Scores

A polygenic risk score (PRS) is a weighted score of independent variants that quantifies the genetic risk to develop a certain trait. As weights for the PRS, we used effect-sizes as found in the summary statistics of the largest GWAS on parental longevity.[15] First, we constructed a PRS using the previously identified longevity variants and tested the association of the PRS

with becoming a cognitively healthy centenarian. We found a significant association of the PRS (OR=1.42, 95% CI=[1.26-1.60], p=6.59x10^{-9}), mainly driven by *APOE* variants (when excluding the *APOE* variants: OR=1.07, 95% CI=[0.96-1.20] and p=0.22) (Figure 5.1 and Table S5). Single-variant association of these variants is available in Table S6. Next, we investigated whether the addition of sub-significant, independent longevity variants increased the association of the PRS with becoming a cognitively healthy centenarian (see section 5.4). The number of additionally included variants to the PRS was based on the association p-value as found in the summary statistics provided by *Timmers et al.*: PRS-8 (p<5x10^{-8}, 19 variants in total), PRS-7 (p<5x10^{-7}, 42 variants in total), PRS-6 (p<5x10^{-6}, 94 variants), PRS-5 (p<5x10^{-5}, 332 variants), PRS-4 (p<0.0005, 1,216 variants), PRS-3 (p<0.005, 3,620 variants), PRS-2 (p<0.05, 8,339 variants) and PRS-1 (p<0.5, 16,926 variants) (Figure 5.1, Table S5, Table S7). For all these PRSs, we tested the difference between cognitively healthy centenarians and population controls. We observed a consistent direction of the effect for all PRSs, with centenarians having on average a higher score than population controls. Including *APOE* variants, we found that the most predictive PRS was the PRS-6, which comprised 96 independent variants (OR=1.44, 95% CI=[1.28-1.61], p=8.39x10^{-10}). Excluding *APOE* variants, the most predictive PRS was the PRS-5, comprising 330 independent variants (OR=1.27, 95% CI=[1.13-1.42], p=4.05x10^{-5}, Figure 5.1, Figure 5.7 and Table S5, Table S7). Single-variant association for all variants is available in Table S8. A more stringent correction for population effects, including 5 additional PCs as covariates, did not change our findings (Table S9). Of note: while controls were a combination of different cohorts, we did not observe cohort-specific associations (Table S10 and Figure 5.11).

5.2.5 Survival analysis

We investigated whether the PRS could predict survival in a subset of the population controls for which follow-up data were available. To investigate the association of the PRS with survival considering *APOE* variants, we performed a survival analysis using the PRS without *APOE* variants with the highest evidence of association in our cohort, *i.e.* PRS-5. We performed a multivariate Cox regression to estimate the association of the PRS-5 with survival while adjusting for age at inclusion, gender, population substructure, and *APOE*-ϵ4 carriership. The PRS-5 was significantly associated with survival in the expected direction (hazard ratio, HR=0.89, 95% CI=[0.80-0.98], p=0.02), *i.e.* having a higher PRS corresponded to reduced mortality. At 50% survival

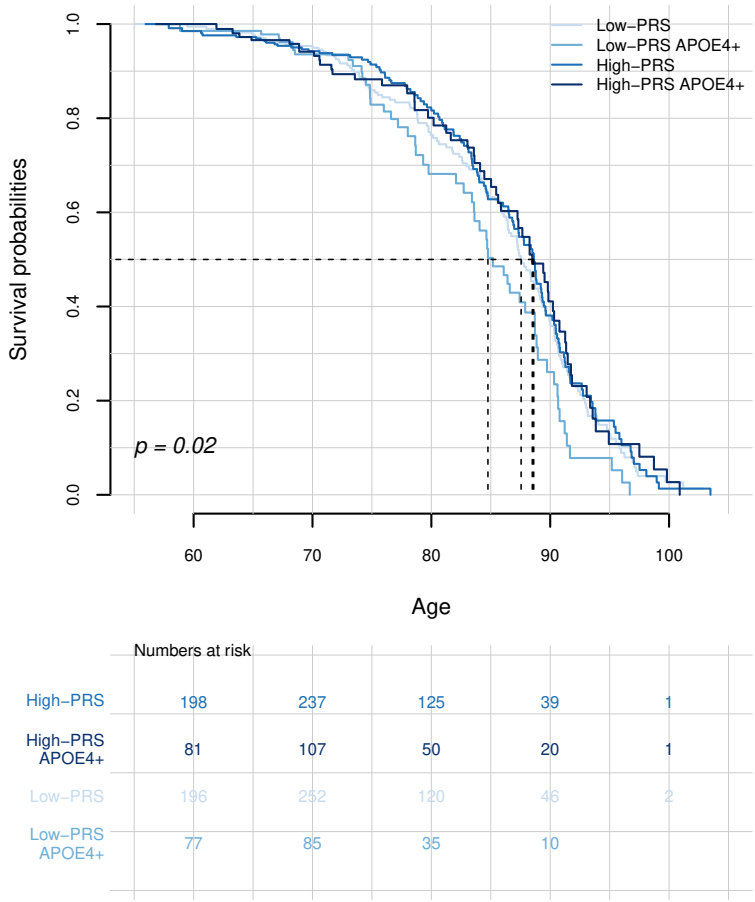


Figure 5.2: **Association of the PRS-5 model without *APOE* on survival, stratified by *APOE*-ϵ4 status.** The top panel shows the survival curves for individuals with *high-PRS* and *low-PRS*, stratified by *APOE*-ϵ4 status (dichotomized in "carriers" of *APOE*-ϵ4 allele and non-carriers) in the population controls for which follow-up data were available (N=1,620). The *p-value* refers to the association of the PRS-5 in the multivariate Cox regression model while adjusting for gender, *APOE*-ϵ4 carriership (dichotomized), and population substructure (PCs 1-5). The lower panel shows the number of samples in each category.

probability (P50), this resulted in 3.86-year difference in survival between individuals with *low-PRS* that were *APOE*-ϵ4 carriers, and those with *high-*

PRS that were not *APOE-ϵ4* carriers (Figure 5.2). We observed that *APOE-ϵ4* carriers with a *low-PRS* had the shortest survival (P50 CI=[0.39-0.65] at age 84.7), followed by non-*APOE-ϵ4* carriers with *low-PRS* (P50 CI=[0.43-0.58] at age 87.5), then, *APOE-ϵ4* carriers with *high-PRS* (P50 CI=[0.38-0.63] at age 88.5) while individuals non-*APOE-ϵ4* carriers with *high-PRS* survived longest (P50 CI=[0.43-0.59] at age 88.6) (Figure 5.2 and Table S11). However, we did not observe a significant interaction effect of PRS and *APOE-ϵ4* status (p=0.27). In line with the known difference in longevity between males and females, gender was significantly associated with survival (HR=1.82 for males compared to females, 95% CI=[1.48-2.26], p=2.72x10^{-8}). A separate analysis in males and females suggested that the PRS was more strongly associated with survival in males than in females (HRM=0.88, 95% CI$_M$=[0.75-1.03] and p_M=0.11; HR$_F$ =0.93, 95% CI$_F$=[0.80-1.05] and p_F =0.24, Figure 5.8). However, we did not find a significant interaction effect between PRS and gender (p=0.60).

5

5.2.6 Functional annotation of PRS

We studied the functional implications of the 330 variants included in PRS-5. First, we linked these variants to 471 unique genes (see section 5.4, Figure 5.9 and Table S12). Then, we looked in the GWAS catalog which variants and associated genes, included in our PRS-5, were previously found to associate with any trait. At the variant-level, of the 330 unique variants, 46 were reported to associate with in total 115 previously analyzed traits, including diseases such as coronary artery disease (CAD, N_{SNP} =13), blood pressure (N_{SNP} =9), and cardiovascular diseases (N_{SNP} =13), but also smoking (N_{SNP} =5) and parental longevity (N_{SNP} =7) (Figure 5.3B). At the gene-level, 300 of the 471 genes in our list were previously associated with lipid metabolism, CAD, neurological traits, and immunological signatures (Figure 5.3C). Next, we performed a gene-set enrichment analysis to explore the biological processes enriched in the 471 PRS-5-associated genes (see section 5.4, also available at `https://snpxplorer.eu.ngrok.io`). We found 48 biological processes significantly enriched after correction for multiple tests (FDR<5%, Table S13), which we reduce to 8 by clustering similar terms together based on semantic similarity measures. These terms pointed towards regulatory and differentiation processes, cellular response to stress, and nervous system development (Figure 5.3A and Table S14). To evaluate the performance of our novel sampling-based method with respect to a traditional gene-set enrichment analysis, we applied the latter to the same 471

genes and we compared the results of both methods. The traditional gene-set enrichment analysis yielded 122 significantly enriched pathways, of which 45 pathways overlap with the 48 significant pathways identified using the sampling-based approach (Table S15). This suggests that our sampling-based approach may be considered conservative compared to traditional gene-set enrichment analyses.

5.2.7 Gene expression of longevity-associated genes

Finally, we studied the expression of the genes linked with the previously identified longevity variants as well as with the PRS-5-associated variants, using a publicly available dataset comprising RNA expression from the hippocampus region in the brain. We compared the RNA-expression in individuals aged 30-65 years (*young*, N=13) as opposed to those aged >80 years (*old*, N=16). We found that 174/432 available genes were differentially expressed after correction for multiple tests (FDR<5%, Figure 5.4 and Table S16): 41 genes were over-expressed in old individuals, while 133 were over-expressed in young individuals.

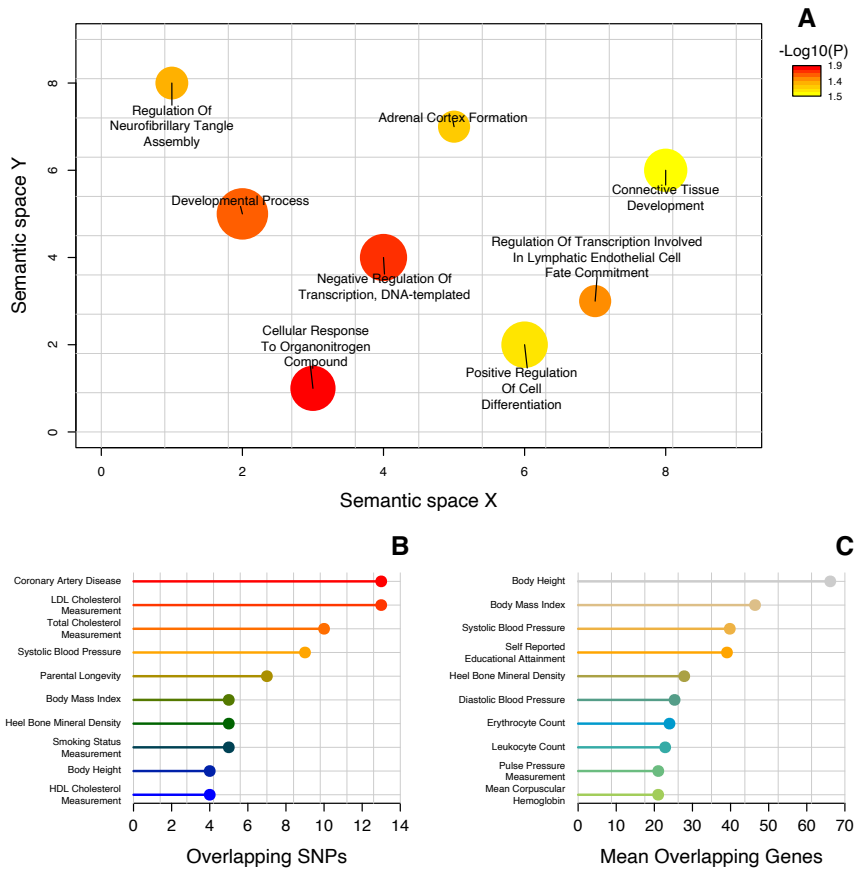


Figure 5.3: **Functional enrichment analysis of variants included in PRS-5 and likely associated genes. A.** Significantly enriched biological processes after merging similar terms based on semantic similarity. Color: encodes significance. Dot size represents the total number of genes associated with that term. **B.** The 10 GWAS catalog terms that have the most overlapping variants in PRS-5. **C.** The top GWAS catalog terms that have the most overlapping genes that associate with the variants in PRS-5.

5.3 Discussion

In this study, we investigated the SNPs underlying extreme human longevity using a sample of cognitively healthy centenarians from the 100-plus Study cohort and a sample of population-matched older-adults. We constructed a polygenic risk score (PRS) comprising 330 variants that was capable of

distinguishing between cognitively healthy centenarians and population controls. This PRS was significantly associated with survival in an independent sample of individuals and may compensate, in part, for the increased mortality risk associated with the *APOE-ϵ4* allele. Using a novel framework, we functionally annotated the variants included in the PRS, which indicated that these were previously associated with cardiometabolic, immunological, oncological, and neurodegenerative conditions. Functional annotation of the genes most likely affected by these variants revealed a significant enrichment for regulatory and differentiation processes, cellular response to stress, and nervous system development.

We constructed a PRS that was associated with becoming a cognitively healthy centenarian and also with prolonged survival across an age continuum, even after excluding the two *APOE* alleles which associated strongest with longevity. Including *APOE* alleles, the PRS comprising 29 previously associated variants significantly associated with becoming a cognitively healthy centenarian, and association statistics only slightly improved upon the addition of variants that sub-significantly associated with longevity. After excluding *APOE* variants, the association of this PRS was not significant, likely due to the different populations and study designs in which the longevity-association of the 29 variants were identified, their low number, and the small effect-sizes. However, the inclusion of sub-significant variants boosted the predictive performance of the PRS, which indicated that these sub-significant variants provide additional distinguishing power, but in aggregate, this is relatively little compared to the strong *APOE* effect. The predictive power of the PRS including 330 variants was highest, and having a *high-PRS* score associated with longer survival in an independent sample of older-adults. We did not identify single-variants driving the increase in distinguishing power effect, such that we assume that all variants contributed similarly. Adding even more variants with lower significance to the PRS decreased association statistics, which eventually stabilized, likely due to random fluctuation of the data.

We explored the relationship between PRS and *APOE-ϵ4* carriership: fully according to expectations, *APOE-ϵ4* carriers with a *low-PRS* had the lowest survival, while as expected, non-*APOE-ϵ4* carriers with a *high-PRS* survived longest, on average 3.86 years longer. Between these extremes, non-*APOE-ϵ4* carriers with *low-PRS* had lower survival compared to *APOE-ϵ4* carriers with a *high-PRS*. This suggests that the variants in the PRS may compensate for the strong disease/mortality risk-increasing effect exerted by the *APOE-ϵ4* allele, however, replication in a large and independent dataset

is needed to confirm this finding. A number of studies described this effect in dementia, and although the results did not strongly replicate across different studies, several variants (*e.g. rs5882* in the *CETP* gene and *rs4934* in the *SERPINA3* gene) were reported to exhibit buffering effects with respect to *APOE-ϵ*4.[27, 28] The majority of the variants included in the best PRS were previously associated with age-related conditions and parental longevity. Given that the variants included were selected from a study on parental longevity, this was not surprising. Functionally, genetic variants were associated with metabolite- and lipid measurements (serum metabolites, total cholesterol, high- and low-density lipoproteins), cardiovascular-related traits (blood pressure, coronary artery diseases, obesity, smoking), neurological conditions (multiple sclerosis, schizophrenia, bipolar disorder) and immunological signatures (IgG glycosylation levels, Crohn's disease, celiac disease). These traits have been associated with longevity either directly, as part of known hallmarks of aging, or indirectly, through their effect on age-related diseases.[1, 8] Likewise, when we investigated the genes associated with the variants in the PRS, we observed an enrichment for mechanisms associated with the aging individual: chronic low-grade inflammation, cellular stress, and a reduced speed of cell-replacement, development, and differentiation.[1]

Recently, increased parental lifespan was associated with a lower PRS of LDL-cholesterol levels, systolic blood pressure, and body mass index.[15] We previously showed that cognitively healthy centenarians have a significantly lower PRS of Alzheimer's Disease (AD) compared to population controls.[29] The overlap between the variants that contribute to the AD-PRS and our best longevity-PRS is limited: apart from *APOE* variants, the longevity-associated variant *rs9665907* is in LD with the known AD-variant *rs11218343* (in/near *SORL1*, R^2=0.39),[30] and the variant *rs6558008* is in low LD with the known AD-variant *rs9331896* (in/near *CLU*, R^2=0.05).[31] This suggests that, in addition to the effect of *APOE* alleles, the *SORL1*- and *CLU*-associated signals may partly overlap in the genetic association of AD and longevity (in opposite directions). Two other studies investigated the relationship between longevity and risk alleles for several age-related diseases: one was able to discriminate between long-lived individuals and controls,[32] while the other did not find significant differences between centenarians and controls.[33] We speculate that the main reason for this discrepancy is that our PRS was constructed based on the association statistics from a well-powered GWAS, which was not available when the previous studies were performed. Additionally, the stricter selection criteria of the centenarians from the 100-plus Study may have contributed to the discriminative power

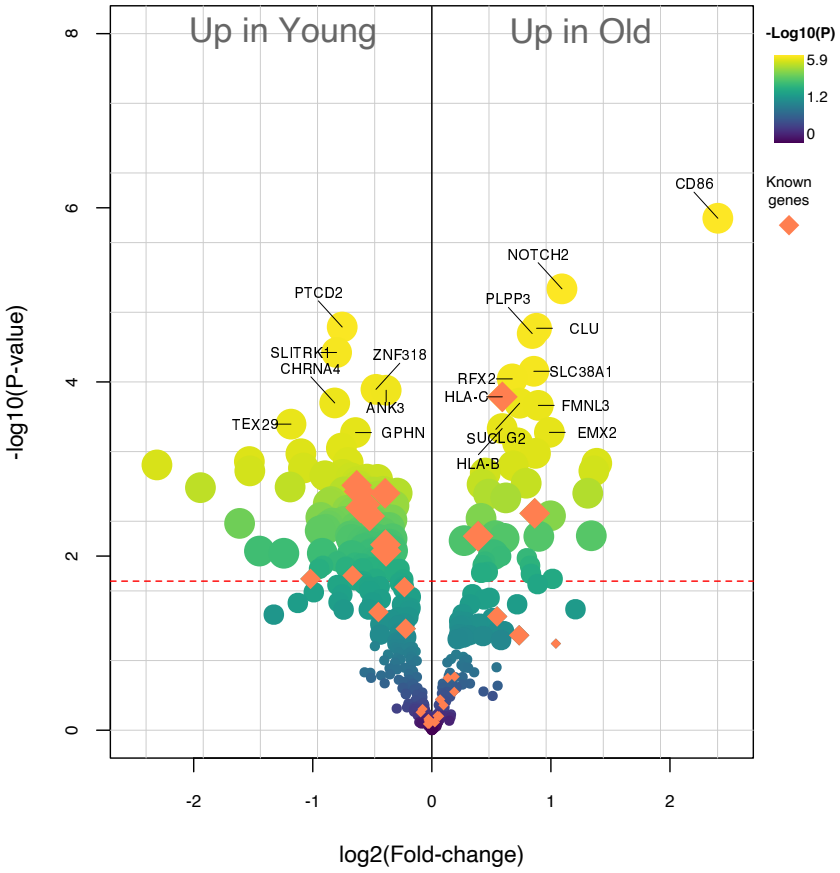


Figure 5.4: **Volcano plot of 432 genes associated with the PRS-5 variants and with the previously identified variants, representing the difference in RNA expression in the hippocampus region of the brain between young individuals (*N*=16) and old individuals (*N*=13) (publicly available dataset GSE11882).** All individuals were reported to be cognitively healthy at the time of death. Orange diamonds refer to 33 genes associated with the previously identified variants. Annotations are shown for the 15 most significant genes as well as for the 3 most significant genes that overlap between the two gene sets.

of the PRS.

Across populations, extreme longevity is known to be more prevalent among females than males, which likely reflects gender differences of environmental exposure, disease predisposition, and genetics.[34] In our study, we found that the effect-size of the PRS on male-survival was larger com-

pared to female-survival, suggesting that males depend more on having advantageous genetic variants to reach extreme ages than females. In the cohort investigated, an important environmental gender-difference is smoking behavior: in accordance with the smoking behaviors in their birth cohort, 76% of the centenarian males had smoked regularly during their lifetime, compared to only 15% of the females.[35] Biological differences may also play a role: estrogens protect females from cardiovascular diseases during their fertile period,[34, 36] produce more vigorous cellular and humoral immune reactions, and are more resistant to infections caused by viruses and other pathogens.[34] From a genetic perspective, impairments in DNA-repair mechanisms become more prevalent with increasing ages, but there are indications that this effect starts a decade earlier in males compared to females.[36] Also, several studies reported that women have longer telomeres compared to males.[36] Together, these studies suggest that females may be more inherently predisposed to live longer than males and that differential exposure to hazardous environments may lead to selective survival of resilient males. Although conclusive evidence that explains the gender-differences in longevity is still lacking, these aspects may in part explain our finding that males are more dependent on an advantageous genetic background to reach extreme ages than females. Note that we did not find a significant interaction effect between PRS and gender, therefore these findings will have to be replicated in a larger cohort.

5.3.1 Strengths and limitations

Linking variants with genes likely affected is difficult: as such, exploiting diverse sources of variant annotations, such as predicted variant consequences, eQTLs, and genomic position is essential to pinpoint the genes likely associated with a variant. We designed a novel framework that allows multiple genes to associate with each variant, in which we consider the annotation-certainties when performing gene-set enrichment analyses. A limitation of our analysis is that our cohort of centenarians is relatively small compared to the sample sizes of previous GWAS. Due to the rarity of this phenotype in the general population, the collection of large cohorts is prohibitive.[20] As a consequence of the limited size, we could not perform exhaustive sex-stratified analyses and thus we cannot exclude that we failed to identify sex-specific associations. Centenarians were compared with a sample of controls combined from different cohorts, yet from the same Dutch population, which may be considered as a strength of our study. While the

inclusion of different cohorts with different inclusion criteria had maximized the available sample size in our study, this could potentially result in confounding effects. However, we assessed that no significant cohort-specific association or population effect affected our results, both at the single-variant and PRS level. We note that our cohort of centenarians was collected in a specific area during a specific time such that location- and period-effects may influence genetic associations. This may in part challenge the replication of the current findings in long-lived individuals from other populations or collected at different times.

5.3.2 Conclusions

We showed that a longevity PRS comprising 330 variants is significantly associated with cognitively healthy aging and with prolonged survival. We found suggestive evidence that the PRS compensates for the deleterious effect of high-impact *APOE-ϵ4* allele and with a novel approach, we functionally annotated the variants in this PRS, showing that many of these variants were previously associated with age-related diseases and with aging-related cellular mechanisms.

5.4 Methods

5.4.1 Study population

As cases, we used a sample of 358 participants from the 100-plus Study cohort.[20] This study includes Dutch-speaking individuals who can provide official evidence for being aged 100 years or older and self-report to be cognitively healthy. As controls, we used (*i*) a sample of 1,779 Dutch older-adults from the Longitudinal Aging Study of Amsterdam (LASA),[37] (*ii*) a sample of 1,206 older-adults with subjective cognitive decline that visited the memory clinic of the Alzheimer center Amsterdam and SCIENCe project, who were labeled cognitively normal after extensive examination,[38] (*iii*) a sample of 40 healthy controls from the Netherlands Brain Bank,[39] (*iv*) a sample of 201 individuals from the twin study,[40] and (*v*) a sample of 86 older-adults from the 100-plus Study (partners of centenarian's children). Individuals with subjective cognitive decline were followed over time in the SCIENCe project, and only individuals who did not convert to mild-cognitive-impairment (MCI) or dementia during follow-up were included in this study. We checked whether the inclusion of controls from cohorts with different inclusion criteria was problematic in terms of cohort-specific associations both at the single-variant level (Table S10) and at the PRS-level (Figure 5.11). The Medical Ethics Committee of the Amsterdam UMC (METC) approved all studies. All participants and/or their legal representatives provided written informed consent for participation in clinical and genetic studies.

5.4.2 Genotyping and imputation procedures

Genetic variants in our cohort were determined by standard genotyping or imputation methods and we applied established quality control methods. All individuals were genotyped using Illumina Global Screening Array (GSAsharedCUSTOM-20018389-A2). We used high-quality genotyping in all individuals (individual call-rate >98%, variant call-rate >98%), individuals with sex mismatches were excluded and departure from Hardy-Weinberg equilibrium was considered significant at $p<1x10^{-6}$. Genotypes were prepared for imputation using available scripts (*HRC-1000G-check-bim.pl*) to compare variant ID, strand, and allele frequencies to the Haplotype Reference Panel (*HRC v1.1*, April 2016).[41] All autosomal variants were submitted to the Sanger imputation server (`https://imputation.sanger.ac.uk`). The server uses MACH to phase data and imputation to the reference panel was performed with PBWT. A total of 3,312 population subjects and 358 centenarians passed quality control. Prior to analysis, we excluded individuals of

non-European ancestry based on 1000Genomes clustering and individuals with a family relationship based on identity-by-descent >0.2.[42] This led to the exclusion of 8 centenarians and 197 controls (non-European) and 7 centenarians and 210 controls (family relations), leaving 2,905 older-adults and 343 cognitively healthy centenarians for the analyses.

5.4.3 Mapping genetic variants to affected genes

We selected 29 genetic variants for which there was evidence of a significant association with longevity from previous GWAS and candidate-gene studies (Table S2), and we linked these variants to their likely affected genes (variant-gene mapping). To do so, we combined annotation from CADD (v1.3),[24, 43] eQTL in blood from GTEx consortium (v8)[25] and positional mapping up to 500 kb from the reported variants (RefSeq build 98).[44] CADD annotation was used to inspect each variant's consequences: in the case of coding variants, we confidently associated the variant with the corresponding gene. For non-coding variants, we first considered possible eQTLs and in case these were not available, we included all genes at increasing distance d from the variant (starting with $d \leq 50kb$, up to $d \leq 500kb$, increasing by $50kb$ until at least one match is found).

5.4.4 Gene-based association

At the gene-level, we combined multiple variants in a gene-based test using MAGMA (v1.06).[26] As genes, we used those that were associated with our variant-gene mapping, and as variants, we used those with minor allele frequency >1% in our population. In MAGMA, we used a flanking window of $2kb$ around each gene and as gene model, and adopted the snp-wise top model (*–gene-model snp-wise=top*), which is most sensible when only a small proportion of SNPs in a gene shows association.[26] Associations were adjusted for population substructure (principal components 1-5) and association p-values were corrected for multiple tests (FDR, correction for the number of genes tested). The number of principal components used as covariates was arbitrarily chosen: given the homogeneous population that we used in this study, we believe this should account for any major population effects. However, we repeated the main associations of the PRS including 5 additional PCs as covariates (Table S9). Before analyses, we explored inflation in MAGMA association statistics: we ran MAGMA with the stated settings for 5,000 randomly selected genes and compared the observed p-value distribution with an expected uniform distribution. The deviation

between the median values of the observed and expected distributions is indicative of test inflation: we found that inflation was 1.1.

5.4.5 Polygenic Risk Scores

We calculated a polygenic risk score (PRS) for each sample in our cohort. As weights for the PRS, we used variant effect sizes (log of odds-ratios) available in the summary statistics of the GWAS on parental longevity.[15] We decided not to use weights from a case-control GWAS as the most recent included our cohort, thus the resulting variant effect-sizes would be biased. Due to the study setting, parental longevity effect-sizes are in general smaller than case-control GWAS of longevity.[14, 15] This would affect the odds-ratios (OR) of our associations, but not the significance, as it would just shift the distribution of the PRS while keeping the same distance between the groups (older-adults and centenarians, in our case). It is then the power of the parental longevity study, due to the large sample size, that determines replicability and predictability of the PRS.[45] Therefore, we believe that using effect-sizes from a parental longevity study has not impacted our findings. The PRSs were Z-standardized and regressed against case-control status (with centenarians as cases and older-adults as controls), correcting for population substructure (PC 1-5). *P*-values were corrected using False Discovery Rate (FDR). Resulting OR can be interpreted as OR-difference per one standard deviation increase in the PRS. We calculated a set of different PRSs: first, using the set of 29 previously identified variants; then, we recursively included in the PRS independent variants that associated sub-significantly with longevity. The inclusion of variants was based on the reported significance in the GWAS summary statistics: PRS-8: $p<5\text{x}10^{-8}$, PRS-7: $p<5\text{x}10^{-7}$, PRS-6: $p<5\text{x}10^{-6}$, PRS-5: $p<5\text{x}10^{-5}$, PRS-4: $p<5\text{x}10^{-4}$, PRS-3: $p<0.005$, PRS-2: $p<0.05$ and PRS-1: $p<0.5$. The selection of independent variants to include in each PRS was performed with LD-based clumping ($R^2<0.001$ within $750kb$ window) using the genotypes of the European samples from the 1000Genome project (phase 3, N=503).[42] Due to their large effect-size, we stratified all PRSs by *APOE* variants, *i.e.* we calculated PRSs with and without *APOE* variants.

5.4.6 Survival analysis

We investigated whether the PRS was predictive for survival in a subset of the older-adults for which follow-up data were available. A total of 1,620 subjects (mean age 62.7±6.4 , 53% female) were eligible for the survival analysis. The age at study inclusion was regarded as *T1*, while the age at

last visit, death, or loss to follow-up was regarded as *T2*, with the survival time calculated as *T2-T1*. The mean follow-up time was 10.4±6.9 years, and at the time of analyses 380 individuals had deceased (23%). Survival analysis was performed implementing left truncation as we anticipated selection bias at old ages, and using the function *Surv(T1, T2, death, type="counting")* as implemented in R-package *survival*. We performed a survival analysis using the (Z-standardized) PRS without *APOE* variants with the highest evidence of association in our cohort (Figure 5.10). Resulting Hazard Ratios (HR) have to be interpreted with respect to 1 unit increase in the PRS. First, we used a multivariate Cox regression to investigate the association of the PRS after correcting for *APOE-ϵ4* status (dichotomized), gender, and population stratification (PC 1-5). For visualization purposes, we split the population into *high-PRS* and *low-PRS* categories based on the median PRS value of the individuals with age <65 years . We then calculated survival differences between the individuals with *low-PRS* and those with *high-PRS* (stratifying for *APOE-ϵ4* status) in a univariate analysis and displayed survival probabilities over age with Kaplan-Meier curves. We calculated differences in years at 50% survival probability between the PRSs. We tested the interaction effect of (*i*) PRS and gender and (*ii*) PRS and *APOE-ϵ4* status on survival by adding an interaction term in the Cox regression model. To evaluate gender-specific effects of the PRS on survival, we repeated the multivariate Cox regression analyses separately in males and females.

5.4.7 Functional annotation of variants comprising the best PRS

We inspected the functional consequences of the variants included in the best PRS model. First, we investigated these variants in the GWAS catalog seeking for previous associations with any trait.[46] Similarly, we looked at whether the genes associated with these variants were previously reported to associate with any trait in the GWAS catalog. To do so, we linked variants to genes as done for the previously identified variants. However, we realized that allowing multiple genes to associate with a variant could result in an enrichment bias, as neighboring genes are often functionally related. To control for this, we implemented sampling techniques (1000 iterations): at each iteration, we (*i*) sampled one gene from the pool of genes associated with each variant (thus allowing only a 1:1 relationship between variants and genes), and (*ii*) looked whether the resulting genes were previously reported in the GWAS catalog. Averaging by the number of iterations, we obtained an unbiased estimation of the overlap of the PRS-associated genes with each

trait in the GWAS catalog. Finally, we investigated the molecular pathways enriched in the PRS-associated genes. Again, we used sampling techniques: at each iteration, we (*i*) sampled one gene from the pool of genes associated with each variant and (*ii*) performed gene-set overlap analysis with the resulting list of genes. Gene-set enrichment analysis was performed with *GOSt* function as implemented in R-package *gprofiler2*, with Biological Processes (GO:BP) as background, excluding electronic annotations and correcting *p*-values using FDR.[47] Finally, we averaged *p*-values for each enriched term over the iterations (*N*=1,000). To reduce the complexity of the resulting enriched biological processes, we exploited the tool *REVIGO*.[48] This tool summarizes enrichment results by removing redundant terms based on a semantic similarity measure, and displays remaining terms in an embedded space via eigenvalue decomposition of the pairwise distance matrix. We chose *Lin* as semantic distance measure and allowed small similarity among terms to be clustered.[49] Last, we compared results from our sampling-based approach with a traditional gene-set enrichment approach, by applying both methods to the full set of genes associated with all variants.

5.4.8 Gene expression of longevity-associated genes

We investigated the expression of the longevity-associated genes using the publicly available dataset *GSE11882*, which comprises RNA-expression from the hippocampus region in the brain. We selected samples reported to be cognitively healthy and aged 30-65 years (*young*, *N*=13) and samples aged 80 years or more (*old*, *N*=16). We performed differential analysis (*old vs. young*) on (*i*) the set of genes associated with the previously reported variants, and (*ii*) the set of PRS-associated genes. Sample selection and differential analysis were performed using the *GEO2R* platform.[50] We corrected *p*-values for multiple tests (FDR) and displayed results with Volcano plot.

5.4.9 Implementation

Quality control of genotype data, population stratification analysis, relatedness analysis and association analysis were performed with *PLINK* (*v2.00a2LM* and *v1.90b4.6*), whereas PRS analysis,[51] functional enrichment analysis, and plots were performed with a mixture of homemade R (*v3.5.2*), bash and Python (*v2.7.14*) scripts. All scripts are available at `https://github.com/TesiNicco/CentenAssoc`. Variant-gene annotation and gene-set enrichment analysis are implemented in a package available at `https://github.com/TesiNicco/AnnotateMe` and can be run at `https://snpxplorer.net`.

5.5 Acknowledgments

The following studies and consortia have contributed to this manuscript. Amsterdam Dementia Cohort (ADC): Research at the Alzheimer center Amsterdam is part of the neurodegeneration research program of Amsterdam Neuroscience. 100-plus Study: we are grateful for the collaborative efforts of all participating centenarians and their family members and/or relatives. Wiesje van der Flier holds the Pasman chair. Longitudinal Aging Study of Amsterdam (LASA): the authors are grateful to all LASA participants, the fieldwork team, and all researchers for their ongoing commitment to the study. Funding: The Alzheimer center Amsterdam is supported by Stichting-Alzheimer-Nederland and Stichting-VUmc fonds. The clinical database structure was developed with funding from Stichting Dioraphte. The SCIENCe project is supported by a research grant from Gieskes-Strijbis fonds. Genotyping of the Dutch case-control samples was performed in the context of EADB (European Alzheimer DNA biobank), funded by the JPco-fuND FP-829-029 (ZonMW project number-733051061). The 100-plus Study was supported by Stichting Alzheimer Nederland (WE09.2014-03), Stichting Diorapthe, horstingstuit foundation, Memorabel (ZonMW project number-733050814), and Stichting VUmc Fonds. Genotyping of the 100-plus study was performed in the context of EADB (European Alzheimer DNA biobank) funded by the JPco-fuND FP-829-029 (ZonMW project number -33051061). LASA is largely supported by a grant from the Netherlands Ministry of Health, Welfare and Sports, Directorate of Long-Term Care. **Conflicts of interest:** all the authors in the study declared no conflict of interest. The funders had no role in the design of the study at any stage.

5.6 Full author list and affiliations

Niccolo' Tesi,[1,2,3] Sven J. van der Lee,[1,2] Marc Hulsman,[1,2,3] Iris E. Jansen,[1,4] Najada Stringa,[5] Natasja M. van Schoor,[5] Martijn Huisman,[5] Philip Scheltens,[1] Marcel J.T. Reinders,[3] Wiesje M. van der Flier,[1,5] and Henne Holstege [1,2,3]

[1] Alzheimer Centre, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[2] Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[3] Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

[4] Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Re-

search, VU, Amsterdam, The Netherlands
[5] Department of Epidemiology and Data Sciences, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

5.7 Supplementary Figures

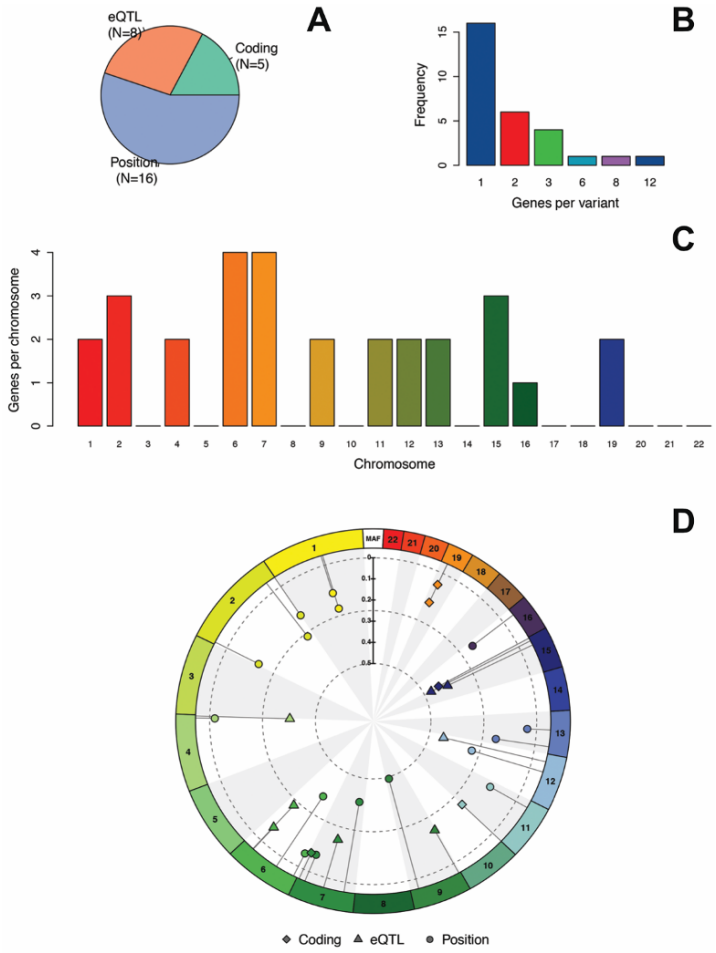


Figure 5.5: **Variant-gene mapping for the 29 previously identified variants. A.** The sources used for the annotation of each variant to the likely affected genes. Coding: variants that code for a change in the amino-acid sequence of the resulting protein; eQTL: variants associated with RNA expression changes in blood from GTEx consortium; Position: variants that are intronic or intergenic. **B.** Histogram of the number of genes associated with each variant. **C**. Distribution of genes across the chromosomes. In total, we mapped 29 variants to 65 unique genes. D. Distribution of the previously identified variants along the genome and each variant's minor allele frequency and annotation.

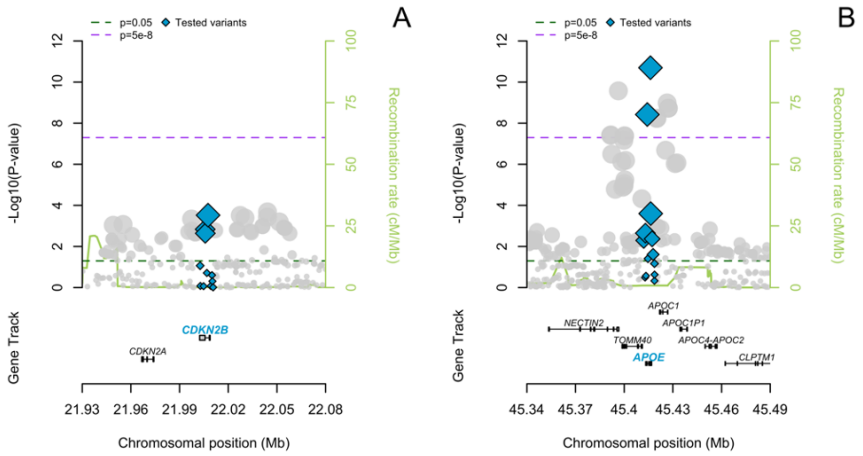


Figure 5.6: **Regional plots of the genetic variants contributing to the gene-based tests that were significantly associated with cognitively healthy aging, in our cohort. A.** Centered on the *CDKN2B* gene. Blue diamonds represent the variants (N=11) that were included in the gene-based test as performed within MAGMA framework. **B.** Centered on the *APOE* gene. Blue diamonds represent the variants (N=13) included in the gene-based test as performed within MAGMA framework. In both figures: genomic positions are plotted on the x-axis (with respect to GRCh37); recombination rates are extracted from HapMap II; RefSeq genes with the largest number of exons are displayed. The dashed blue line indicates the threshold commonly adopted in GWAS for genome-wide significance.

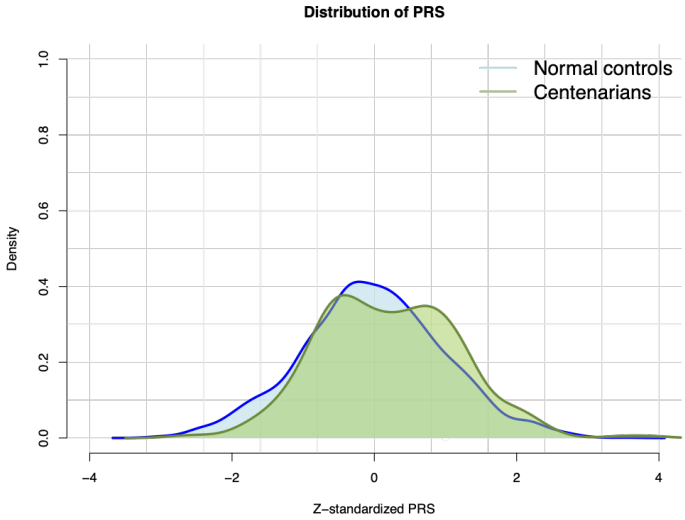


Figure 5.7: **Density distribution of the PRS-5 without APOE variants in population controls and cognitively healthy centenarians.** The figure shows the distribution density of the PRS-5 including 330 variants (excluding *APOE* variants). PRS was Z-standardized.

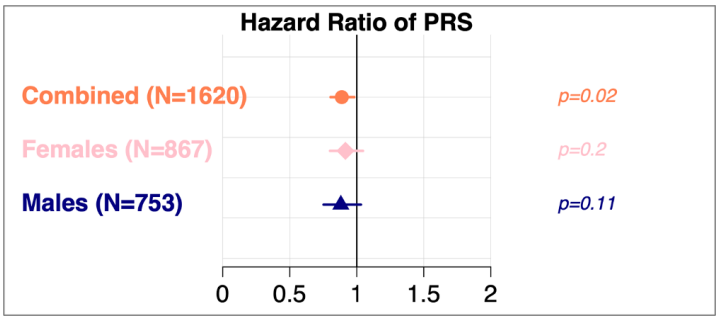


Figure 5.8: **Effect of PRS-5 on survival.** Forest plot of the hazard ratios for PRS-5 in males (N=753), females (N=867), and combined (N=1620). The p-values refer to the association of PRS-5 in the multivariate Cox regression model while adjusting for gender (for the combined analysis only), *APOE*-ϵ4 carriership (dichotomized) and population substructure (PCs 1-5).

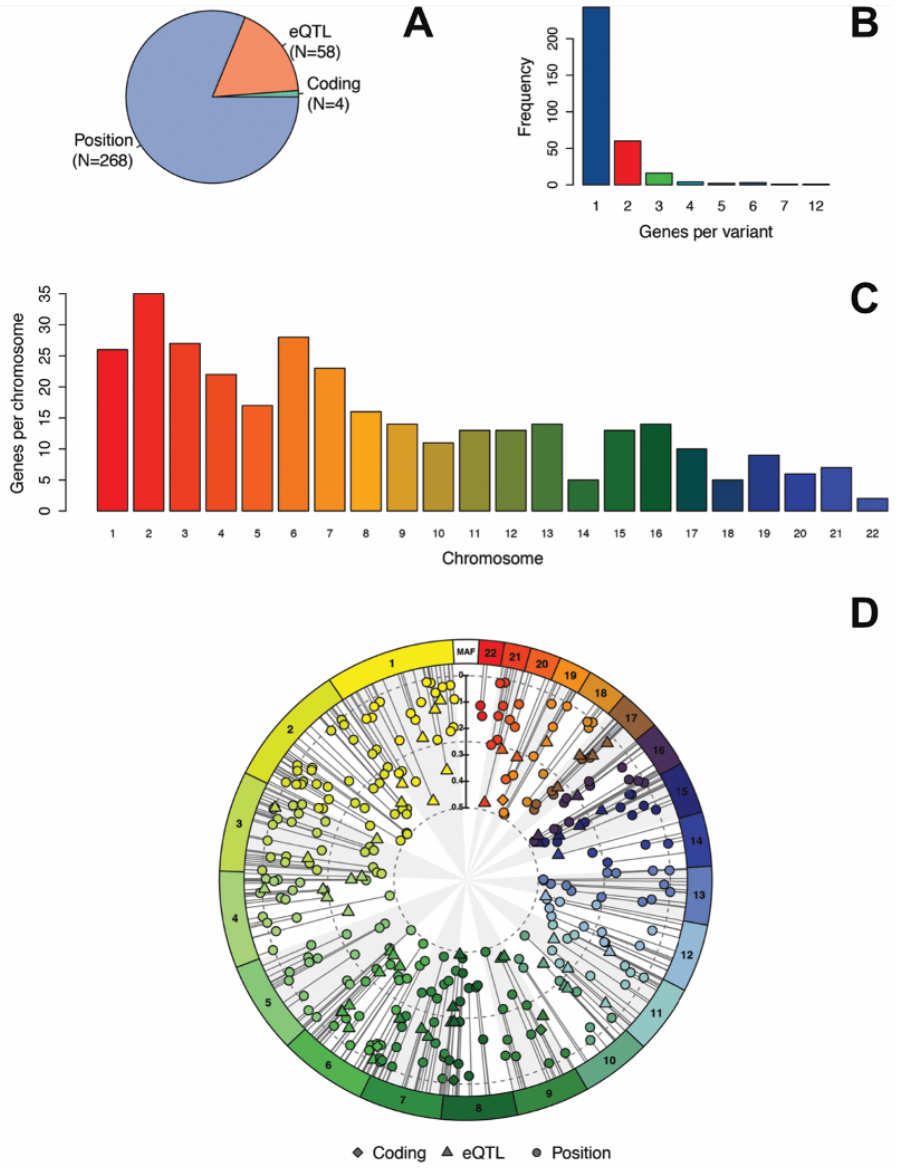


Figure 5.9: **Variant-gene mapping for the variants included in PRS-5. A.** The resource used for the annotation of each variant. *Coding*: variants that code for a change in the amino-acid sequence of the resulting protein; *eQTL*: variants associated with RNA expression changes in blood from GTEx consortium; *Position*: variants that are intronic or intergenic. **B.** Histogram of the number of genes associated with each variant. **C.** Distribution of genes across the chromosomes. In total, we mapped 330 variants to 471 unique genes. **D.** Distribution of variants included in PRS-5 along the genome and each variant's minor allele frequency and annotation source.

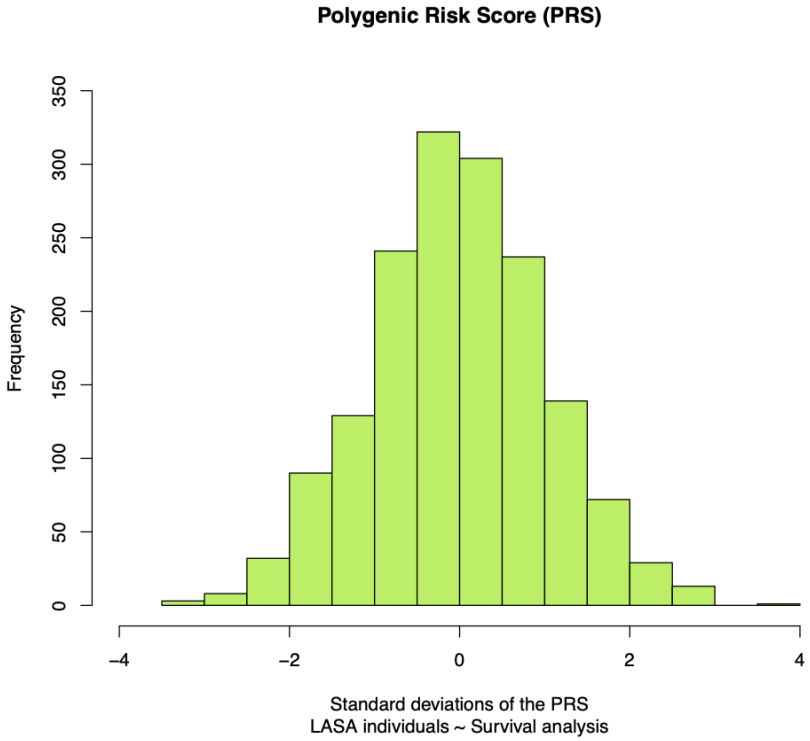


Figure 5.10: **Distribution of PRS-5 in the subset of population controls used in the survival analysis.** The figure shows the distribution of PRS-5 in the subset of controls (N=1630) from the Longitudinal Aging Study of Amsterdam for which follow-up data was available, that we used for the survival analysis. PRS were Z-standardized (μ=0, σ=1).

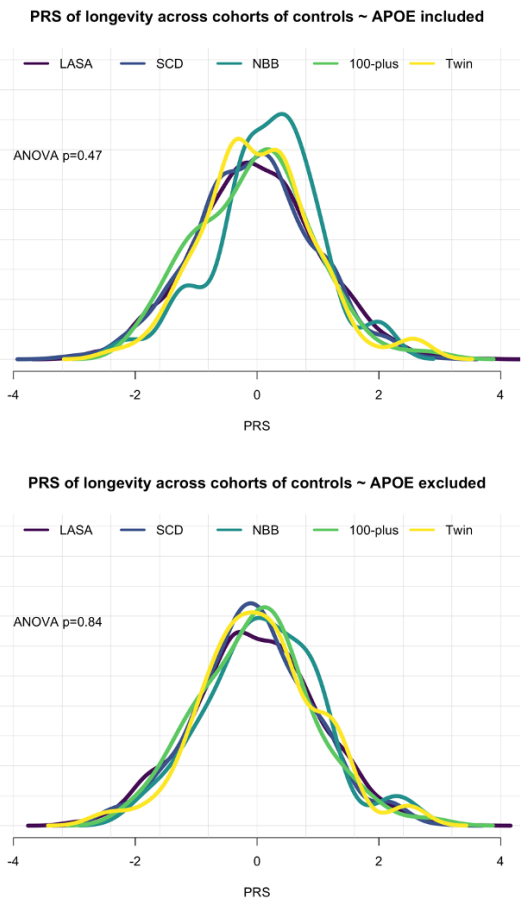


Figure 5.11: **Figure S7: Distribution of PRS-5 comprising 330 genetic variants across the different cohorts of older-adults that were used as controls.** The figure shows the distribution of the PRS-5 including (plot above) and excluding (plot below) *APOE* variants in individuals (*i*) from the Longitudinal Aging Study of Amsterdam (LASA, N=1,648), (*ii*) with subjective cognitive decline that were labeled cognitively healthy after cognitive examination (SCD, N=1,038), (*iii*) cognitively healthy from the Netherlands Brain Bank (NBB, N=37), (*iv*) cognitively healthy from the twin study of Amsterdam (Twin, N=100) and (*v*) the partners of the centenarian's children from the 100-plus Study (100-plus, N=82). We tested for differences in PRS across all cohorts with an anova test, and report the relative p-values in each plot.

5.8 Supplementary Tables

Supplementary Tables can be accessed by scanning the following code or accessing the journal's website here.

References

[1] Linda Partridge, Joris Deelen, and P. Eline Slagboom. "Facing up to the global challenges of ageing". In: *Nature* 561.7721 (Sept. 2018), pp. 45–56. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0457-8.

[2] Angela R. Brooks-Wilson. "Genetics of healthy aging and longevity". In: *Human Genetics* 132.12 (Dec. 2013), pp. 1323–1338. ISSN: 1432-1203. DOI: 10.1007/s00439-013-1342-z.

[3] J. Graham Ruby et al. "Estimates of the Heritability of Human Longevity Are Substantially Inflated due to Assortative Mating". In: *Genetics* 210.3 (Nov. 2018), pp. 1109–1124. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.118.301613.

[4] Joanna Kaplanis et al. "Quantitative analysis of population-scale family trees with millions of relatives". In: *Science* 360.6385 (Apr. 13, 2018), pp. 171–175. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aam9309.

[5] Paola Sebastiani and Thomas T. Perls. "The genetics of extreme longevity: lessons from the new England centenarian study". In: *Frontiers in Genetics* 3 (2012), p. 277. ISSN: 1664-8021. DOI: 10.3389/fgene.2012.00277.

[6] Paolo Garagnani et al. "Centenarians as super-controls to assess the biological relevance of genetic risk factors for common age-related diseases: a proof of principle on type 2 diabetes". In: *Aging* 5.5 (May 2013), pp. 373–385. ISSN: 1945-4589. DOI: 10.18632/aging.100562.

[7] Niccolò Tesi et al. "Centenarian controls increase variant effect sizes by an average twofold in an extreme case–extreme control analysis of Alzheimer's disease". In: *European Journal of Human Genetics* (Sept. 2018). ISSN: 1018-4813, 1476-5438. DOI: 10.1038/s41431-018-0273-5.

[8] David Melzer, Luke C. Pilling, and Luigi Ferrucci. "The genetics of human ageing". In: *Nature Reviews Genetics* (Nov. 2019). ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-019-0183-6.

[9] Param Priya Singh et al. "The Genetics of Aging: A Vertebrate Perspective". In: *Cell* 177.1 (Mar. 2019), pp. 200–220. ISSN: 00928674. DOI: 10.1016/j.cell.2019.02.038.

[10] Vivian Tam et al. "Benefits and limitations of genome-wide association studies". In: *Nature Reviews Genetics* 20.8 (Aug. 2019), pp. 467–484. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-019-0127-1.

[11] Linda Broer et al. "GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy". In: *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 70.1 (Jan. 2015), pp. 110–118. ISSN: 1758-535X. DOI: 10.1093/gerona/glu166.

[12] Paola Sebastiani et al. "Four Genome-Wide Association Studies Identify New Extreme Longevity Variants". In: *The Journals of Gerontology: Series A* 72.11 (Oct. 2017), pp. 1453–1464. ISSN: 1079-5006, 1758-535X. DOI: 10.1093/gerona/glx027.

[13] Kristen Fortney et al. "Genome-Wide Scan Informed by Age-Related Disease Identifies Loci for Exceptional Human Longevity". In: *PLOS Genetics* 11.12 (Dec. 2015). Ed. by Hao Li, e1005728. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1005728.

[14] Joris Deelen et al. “A meta-analysis of genome-wide association studies identifies multiple longevity genes”. In: *Nature Communications* 10.1 (Dec. 2019). ISSN: 2041-1723. DOI: 10.1038/s41467-019-11558-2.

[15] Paul RHJ Timmers et al. “Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances”. In: *eLife* 8 (Jan. 2019). ISSN: 2050-084X. DOI: 10.7554/eLife.39856.

[16] Yi Zeng et al. “Novel loci and pathways significantly associated with longevity”. In: *Scientific Reports* 6.1 (Aug. 2016). ISSN: 2045-2322. DOI: 10.1038/srep21243.

[17] Nuria Garatachea et al. “ApoE gene and exceptional longevity: Insights from three independent cohorts”. In: *Experimental Gerontology* 53 (May 2014), pp. 16–23. ISSN: 05315565. DOI: 10.1016/j.exger.2014.02.004.

[18] Cristina Giuliani, Paolo Garagnani, and Claudio Franceschi. “Genetics of Human Longevity Within an Eco-Evolutionary Nature-Nurture Framework”. In: *Circulation Research* 123.7 (Sept. 2018), pp. 745–772. ISSN: 0009-7330, 1524-4571. DOI: 10.1161/CIRCRESAHA.118.312562.

[19] Cristina Giuliani et al. “Centenarians as extreme phenotypes: An ecological perspective to get insight into the relationship between the genetics of longevity and age-associated diseases”. In: *Mechanisms of Ageing and Development* 165 (July 2017), pp. 195–201. ISSN: 00476374. DOI: 10.1016/j.mad.2017.02.007.

[20] Henne Holstege et al. “The 100-plus Study of Dutch cognitively healthy centenarians: rationale, design and cohort description”. In: (Apr. 2018). DOI: 10.1101/295287.

[21] Thomas Perls. “Dementia-free centenarians”. In: *Experimental Gerontology* 39.11 (Nov. 2004), pp. 1587–1593. ISSN: 05315565. DOI: 10.1016/j.exger.2004.08.015.

[22] Nina Beker et al. “Longitudinal Maintenance of Cognitive Health in Centenarians in the 100-plus Study”. In: *JAMA Network Open* 3.2 (Feb. 2020), e200094. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2020.0094.

[23] Nina Beker et al. “Neuropsychological Test Performance of Cognitively Healthy Centenarians: Normative Data From the Dutch 100-Plus Study: COGNITIVE PERFORMANCE IN CENTENARIANS”. In: *Journal of the American Geriatrics Society* 67.4 (Apr. 2019), pp. 759–767. ISSN: 00028614. DOI: 10.1111/jgs.15729.

[24] Philipp Rentzsch et al. “CADD: predicting the deleteriousness of variants throughout the human genome”. In: *Nucleic Acids Research* 47 (D1 Jan. 2019), pp. D886–D894. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gky1016.

[25] GTEx Consortium. “The Genotype-Tissue Expression (GTEx) project”. In: *Nature Genetics* 45.6 (June 2013), pp. 580–585. ISSN: 1546-1718. DOI: 10.1038/ng.2653.

[26] Christiaan A. de Leeuw et al. “MAGMA: Generalized Gene-Set Analysis of GWAS Data”. In: *PLOS Computational Biology* 11.4 (Apr. 2015). Ed. by Hua Tang, e1004219. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004219.

[27] Erin Elizabeth Sundermann et al. "Cholesteryl ester transfer protein genotype modifies the effect of apolipoprotein $\epsilon 4$ on memory decline in older adults". In: *Neurobiology of Aging* 41 (May 2016), 200.e7–200.e12. ISSN: 01974580. DOI: 10.1016/j.neurobiolaging.2016.02.006.

[28] Aamira J. Huq et al. "Genetic resilience to Alzheimer's disease in *APOE* $\epsilon 4$ homozygotes: A systematic review". In: *Alzheimer's & Dementia* 15.12 (Dec. 2019), pp. 1612–1623. ISSN: 15525260. DOI: 10.1016/j.jalz.2019.05.011.

[29] Niccolò Tesi et al. "Immune response and endocytosis pathways are associated with the resilience against Alzheimer's disease". In: *Translational Psychiatry* 10.1 (Dec. 2020), p. 332. ISSN: 2158-3188. DOI: 10.1038/s41398-020-01018-7.

[30] Ekaterina Rogaeva et al. "The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease". In: *Nature Genetics* 39.2 (Feb. 2007), pp. 168–177. ISSN: 1061-4036. DOI: 10.1038/ng1943.

[31] Denise Harold et al. "Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease". In: *Nature Genetics* 41.10 (Oct. 2009), pp. 1088–1093. ISSN: 1546-1718. DOI: 10.1038/ng.440.

[32] Paola Sebastiani et al. "Genetic Signatures of Exceptional Longevity in Humans". In: *PLoS ONE* 7.1 (Jan. 2012). Ed. by Greg Gibson, e29848. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0029848.

[33] Marian Beekman et al. "Genome-wide association study (GWAS)-identified disease risk alleles do not compromise human longevity". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.42 (Oct. 2010), pp. 18046–18049. ISSN: 1091-6490. DOI: 10.1073/pnas.1003540107.

[34] Rita Ostan et al. "Gender, aging and longevity in humans: an update of an intriguing/neglected scenario paving the way to a gender-specific medicine". In: *Clinical Science* 130.19 (Oct. 1, 2016), pp. 1711–1725. ISSN: 0143-5221, 1470-8736. DOI: 10.1042/CS20160004.

[35] Frans Van Poppel et al. "Mortality decline and reproductive change during the Dutch demographic transition: Revisiting a traditional debate with new data". In: *Demographic Research* 27 (Aug. 23, 2012), pp. 299–338. ISSN: 1435-9871. DOI: 10.4054/DemRes.2012.27.11.

[36] Kathleen E Fischer and Nicole C Riddle. "Sex Differences in Aging: Genomic Instability". In: *The Journals of Gerontology: Series A* 73.2 (Jan. 16, 2018), pp. 166–174. ISSN: 1079-5006, 1758-535X. DOI: 10.1093/gerona/glx105.

[37] Emiel O. Hoogendijk et al. "The Longitudinal Aging Study Amsterdam: cohort update 2016 and major findings". In: *European Journal of Epidemiology* 31.9 (Sept. 2016), pp. 927–945. ISSN: 0393-2990, 1573-7284. DOI: 10.1007/s10654-016-0192-0.

[38] Wiesje M. van der Flier and Philip Scheltens. "Amsterdam Dementia Cohort: Performing Research to Optimize Care". In: *Journal of Alzheimer's Disease* 62.3 (Mar. 2018). Ed. by George Perry, Jesus Avila, and Xiongwei Zhu, pp. 1091–1111. ISSN: 13872877, 18758908. DOI: 10.3233/JAD-170850.

[39] Marleen C. Rademaker, Geertje M. de Lange, and Saskia J.M.C. Palmen. "The Netherlands Brain Bank for Psychiatry". In: *Handbook of Clinical Neurology*. Vol. 150. Elsevier, 2018, pp. 3–16. ISBN: 978-0-444-63639-3. DOI: 10.1016/B978-0-444-63639-3.00001-3.

[40] Gonneke Willemsen et al. "The Netherlands Twin Register Biobank: A Resource for Genetic Epidemiological Studies". In: *Twin Research and Human Genetics* 13.3 (June 2010), pp. 231–245. ISSN: 1832-4274, 1839-2628. DOI: 10.1375/twin.13.3.231.

[41] Shane McCarthy et al. "A reference panel of 64,976 haplotypes for genotype imputation". In: *Nature Genetics* 48.10 (Oct. 2016), pp. 1279–1283. ISSN: 1546-1718. DOI: 10.1038/ng.3643.

[42] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (Oct. 2015), pp. 68–74. ISSN: 1476-4687. DOI: 10.1038/nature15393.

[43] Martin Kircher et al. "A general framework for estimating the relative pathogenicity of human genetic variants". In: *Nature Genetics* 46.3 (Mar. 2014), pp. 310–315. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.2892.

[44] Nuala A. O'Leary et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic Acids Research* 44 (D1 Jan. 2016), pp. D733–745. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1189.

[45] Frank Dudbridge. "Power and Predictive Accuracy of Polygenic Risk Scores". In: *PLoS Genetics* 9.3 (Mar. 2013). Ed. by Naomi R. Wray, e1003348. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003348.

[46] Annalisa Buniello et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". In: *Nucleic Acids Research* 47 (D1 Jan. 2019), pp. D1005–D1012. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gky1120.

[47] Uku Raudvere et al. "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)". In: *Nucleic Acids Research* 47 (W1 July 2019), W191–W198. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz369.

[48] Fran Supek et al. "REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms". In: *PLoS ONE* 6.7 (July 2011). Ed. by Cynthia Gibas, e21800. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021800.

[49] Bridget T. McInnes and Ted Pedersen. "Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text". In: *Journal of Biomedical Informatics* 46.6 (Dec. 2013), pp. 1116–1124. ISSN: 15320464. DOI: 10.1016/j.jbi.2013.08.008.

[50] Tanya Barrett et al. "NCBI GEO: archive for functional genomics data sets—update". In: *Nucleic Acids Research* 41 (D1 Nov. 2012), pp. D991–D995. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gks1193.

[51] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *American Journal of Human Genetics* 81.3 (Sept. 2007), pp. 559–575. ISSN: 0002-9297. DOI: 10.1086/519795.

6. A large GWAS of Alzheimer's disease

Common variants in Alzheimer's disease and risk stratification by polygenic risk scores

Itziar de Rojas,* Sonia Moreno-Grau,* Niccolò Tesi,* Benjamin Grenier-Boley,* Victor Andrade,* Iris Jansen,* Nancy L. Pedersen, Najada Stringa, Anna Zettergren, Isabel Hernández *et al.*

* Authors contributed equally

This chapter was published in *Nature Communications*
https://www.nature.com/articles/s41467-021-22491-8

Abstract

Genetic discoveries of Alzheimer's disease are the drivers of our understanding, and together with polygenetic risk stratification can contribute towards planning of feasible and efficient preventive and curative clinical trials. We first perform a large genetic association study by merging all available case-control datasets and by-proxy study results (discovery n=409,435 and validation size n=58,190). Here, we add six variants associated with Alzheimer's disease risk (near *APP, CHRNE, PRKD3/NDUFAF7, PLCG2* and two exonic variants in the *SHARPIN* gene). Assessment of the polygenic risk score and stratifying by *APOE* reveal a 4 to 5.5 years difference in median age at onset of Alzheimer's disease patients in *APOE* ϵ4 carriers. Because of this study, the underlying mechanisms of *APP* can be studied to refine the amyloid cascade and the polygenic risk score provides a tool to select individuals at high risk of Alzheimer's disease.

6.1 Background

Thus far, multiple loci associated with Alzheimer's disease (AD) have been described next to causal mutations in two subunits of γ-secretases, membrane-embedded aspartyl complexes (*PSEN1, PSEN2* genes), and the gene encoding one target protein of these proteases, the amyloid precursor protein gene (*APP*). The most prominent locus, *APOE*, was detected almost 30 years ago using linkage techniques.[1] In addition, genome-wide association studies (GWAS) of AD case-control datasets and by-proxy AD case-control studies have identified 30 genomic loci that modify the risk of AD.[2, 3, 4, 5, 6] These signals account for 31% of the genetic variance of AD, leaving most of the genetic risk as yet uncharacterized.[7] Further disentangling the genetic constellation of common genetic variations underlying AD can drive our biological insights of AD and can point toward novel drug targets. There are over 50 million people living with dementia and the global cost of dementia is well above 1 trillion US$.[8] This means there is a medical and economical urgency to efficiently test interventions that are under development. Therefore, to increase power and reduce duration of trials, pre-symptomatic patients that are at high genetic risk of disease are increasingly developed.[9] However, only carriers of causal mutations (*APP, PSEN1, PSEN2*) and the *APOE* $\epsilon 4$ allele are considered high risk, while other common and rare genetic variants are ignored.[10] Despite that, the combined effects of all currently known variants in a polygenic risk score (PRS) is associated with the conversion of mild cognitive impairment (MCI) to AD,[11, 12], the neuropathological hallmarks of AD, age at onset (AAO) of disease [13, 14, 15, 16] and lifetime risk of AD.[17] Here, we aimed to comprehend and expand the knowledge of the genetic landscape underlying AD and provide additional evidence that a PRS of variants can be a robust tool to select high risk individuals with an earlier age at onset. We first performed a meta-GWAS integrating all currently published GWAS case-control data, by-proxy case-control data, and the data from the Genome Research at Fundació ACE (GR@ACE) study.[18] We confirmed the novel observed associations in a large independent replication study. Then, we constructed an update of the PRS and tested whether the effects of the PRS were influenced by diagnostic certainty, sex and AAO groups. Lastly, we tested whether the PRS could be used to identify individuals at the highest odds of having AD and we compared age at onset of the AD cases.

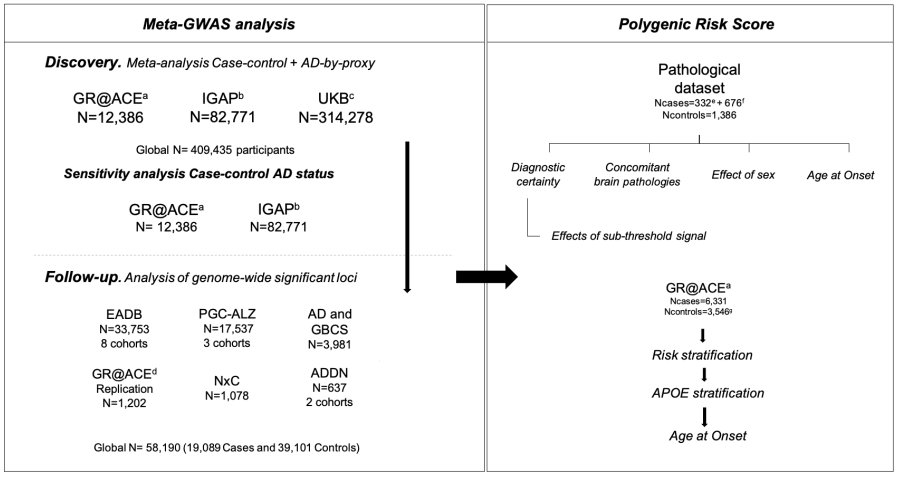


Figure 6.1: **Flow chart of analysis steps.** Discovery meta-analysis in GR@ACE, IGAP stage 1+2 and UKBiobank followed by a replication in 16 independent cohorts. The genome-wide significant signals found in meta-GWAS were used to perform a Polygenic Risk Score in a clinical and pathological AD dataset. See Supplementary Methods for more information about the cohorts included and methods to the PRS generation. ^{a}, extended dataset (S.Moreno-Grau *et al.* 2019)[18]; ^{b}, stage I + stage II (Kunkle *et al.* 2019)[19]; ^{c}, by proxy AD: meta-analysis of maternal and paternal history of dementia (Marioni *et al.* 2018)[20]; ^{d}, extra and independent GR@ACE dataset incorporated only for replication purposes; ^{e}, pathologically confirmed AD cases; ^{f}, AD cases diagnosed based on clinical criteria; ^{g}, controls participants aged 55 years and younger. *N*=, total of individuals within specified data.

6.2 Results

6.2.1 Meta-GWAS of AD

We combined data from three AD GWASs: the summary statistics calculated from the GR@ACE case-control study (6,331 AD cases and 6,055 controls),[18] the IGAP case-control study (up to 30,344 AD cases and 52,427 controls) [19] and the UKB AD-by-proxy case-control study (27,696 cases of maternal AD with 260,980 controls and 14,338 cases of paternal AD with 245,941 controls, Figure 6.1, Supplementary Table 1).[20] Although we observed inflation in the resulting summary statistics (λ=1.08; see Figure 6.7d), it was not driven by an un-modeled population structure (LD score regression intercept=1.04). The full details of the studies are described in the supplementary methods. After study-specific variant filtering and quality-control procedures, we performed a fixed-effects inverse-variance-weighted meta-analysis on the

summary statistics of the three studies.[21] Using this strategy, we identified a genome-wide significant (GWS) association (p<5x10^{-8}) for 36 independent genetic variants in 35 genomic regions (the *APOE* region contains signals for ϵ4 and ϵ2). As a sensitivity analysis, we removed the AD-by-proxy study and compared the resulted effect estimates with and without this dataset. We found a high correlation between the effect estimates from the case-control and by-proxy approaches for the significant loci (R^2 = 0.994, p = 8.1x10^{-37}; Figure 6.7e). Four genomic regions were not previously associated with AD (see Manhattan Plot, Figure 6.2).

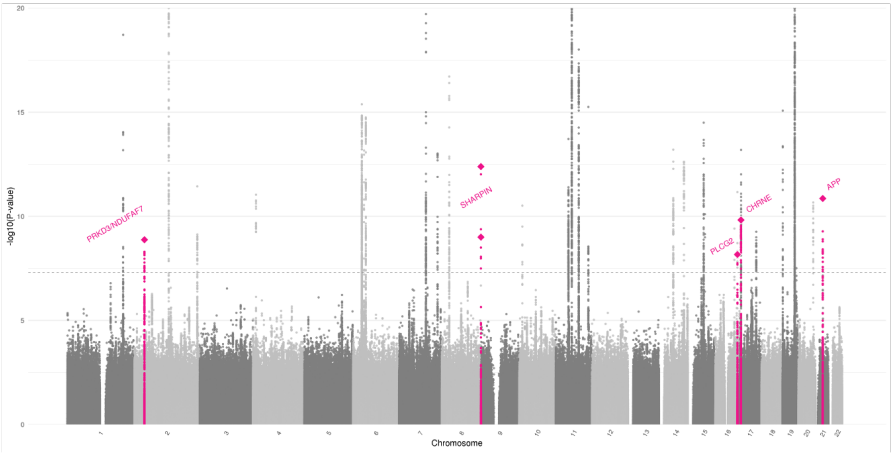


Figure 6.2: **Manhattan plot of the overall GWAS meta-analysis for AD risk (N = 467,623)**. Genome-wide associations with Alzheimer's disease highlighting the novel loci associated with AD (*PRKD3/NDUFAF7, SHARPIN, CHRNE, PLCG2* and *APP*).

Next, we aimed at replicating the associated loci in 16 cohorts (19,087 AD cases and 39,101 controls in total), many of them collected and analyzed by the European Alzheimer's Disease Biobank (JPND-EADB) project. We tested all variants with suggestive association (p<10^{-5}) located within a 200Kb region from the sentinel SNP. Overall, 384 variants were tested in the replication datasets (Supplementary Table 2). Discovery and replication were combined, and we identified novel associations in six variants comprising five genomic loci annotated using FUMA (Table 6.1, Figure 6.3D-F, Figure 6.8 and Supplementary Results).[22] In *APP*, we identified a common (MAF = 0.46) intronic variant associated with a reduced risk of AD (*rs2154481*, *OR* = 0.95 [0.93-0.96], p= 9.3x10^{-10}, Figure 6.3F). In *SHARPIN* (SHANK As-

sociated RH Domain Interactor) gene, we found two missense mutations (*rs34173062*/p.Ser17Phe and *rs34674752*/p.Pro294Ser) that are in linkage equilibrium (R^2=1.3x10^{-6}, *D*'=0.014, p = 0.96). Both missense variants increased AD risk (p.Ser17Phe, MAF = 0.085, *OR* = 1.14 [1.10-1.18], p = 9.6x10^{-13} and p.Pro294Ser, MAF = 0.052, *OR* = 1.13 [1.09-1.18], p = 1.0x10^{-9}, Figure 6.3A-B). A variant close to the genes *PRKD3* and *NDUFAF7* (*rs876461*, MAF = 0.143) emerged as the most significant variant in the region after the combined analysis (*OR* = 1.07 [1.05-1.09], p = 1.3x10^{-9}, Figure 6.3C). In the 3'-UTR region of *CHRNE* (Cholinergic Receptor Nicotinic Epsilon Subunit), *rs72835061* (MAF = 0.085) was associated with a 1.09-fold increased risk of AD (*95% CI* [1.06-1.11], p = 1.5x10^{-10}, Figure 6.3E). Our analysis also strengthened the evidence of association with AD for three additional genomic loci including a novel association with a variant in *PLCG2* (*rs3935877*, MAF= 0.13, *OR* = 0.92 [0.90-0.95], p = 6.9x10^{-9}, Figure 6.3D), and confirmed another common variant in *PLCG2*, a stop gain mutation in *IL34* and a variant near *HS3ST1* (Table 6.1, Figure 6.9 and Supplementary Tables 2-3). We were not able to replicate two loci (*ELK2AP* and *SPPL2A* regions) that showed suggestive association with AD (p<1x10^{-7} in discovery).

6.2.2 Polygenic Risk Scores

In order to assess the robustness and combined effect of the new genetic landscape of AD (Figure 6.4, Supplementary Table 4), we constructed a weighted PRS based on the 39 genetic variants (excluding *APOE* genotypes) that showed GWS evidence of association with AD (see section 6.4, Figure 6.5 and Supplementary Table 5). We tested if the association of the PRS with AD is independent of clinically important factors that are considered in the selection of individuals for clinical trials. First, we showed that the association of the PRS with clinically diagnosed AD cases is similar to the association with pathologically confirmed AD (*OR* = 1.30 *vs.* 1.38, per 1-SD increase in the PRS). In this setting, adding variants below the GWS threshold did not lead to a more significant association of the PRS with AD (Figure 6.5A). Next, we tested whether the PRS was associated with AD in the presence of concomitant brain pathologies (besides AD). Among our autopsy-confirmed AD patients (n=332), 84% had at least one concomitant pathology, and the PRS was associated with AD in the presence of all tested concomitant pathologies (Figure 6.5B). Moreover, the patients often had more than one concomitant pathology (48.8%), but no difference was observed in the effect estimate of the PRS when more than one pathology was present (Figure 6.5B). Last,

Table 6.1: **Results for the AD loci selected for follow-up**

							Discovery meta-analysis		Follow-up datasets		Overall	
Chr	Pos	SNP	Closest gene	A1	A2	Freq A1	OR [95% CI]	P	OR [95% CI]	P	OR [95% CI]	P
2	37515958	*rs876461*	*PRKD3/NDUFAF7*	A	G	0.143	1.07 [1.04-1.09]	$9.14x10^{-7}$	1.08 [1.04-1.13]	$3.07x10^{-4}$	1.07 [1.05-1.09]	$1.34x10^{-9}$
8	145154222	*rs34674752*	*SHARPIN*	A	G	0.052	1.11 [1.06-1.16]	$4.02x10^{-6}$	1.20 [1.10-1.31]	$1.65x10^{-5}$	1.13 [1.09-1.18]	$1.00x10^{-9}$
8	145158607	*rs34173062*	*SHARPIN*	A	G	0.085	1.16 [1.11-1.21]	$1.33x10^{-11}$	1.09 [1.02-1.17]	$7.35x10^{-3}$	1.14 [1.10-1.18]	$9.62x10^{-13}$
16	81900853	*rs3935877*	*PLCG2*	C	T	0.868	0.92 [0.90-0.95]	$1.12x10^{-7}$	0.92 [0.85-0.99]	$1.96x10^{-2}$	0.92 [0.90-0.95]	$6.85x10^{-9}$
17	4805437	*rs72835061*	*CHRNE*	A	C	0.085	1.09 [1.06-1.12]	$3.92x10^{-9}$	1.07 [1.02-1.12]	$7.83x10^{-3}$	1.09 [1.06-1.11]	$1.51x10^{-10}$
21	27473875	*rs2154481*	*APP*	C	T	0.483	0.95 [0.93-0.96]	$9.26x10^{-10}$	0.96 [0.93-0.99]	$3.31x10^{-3}$	0.95 [0.94-0.96]	$1.39x10^{-11}$
Previously reported genome-wide significant hits replicating in the follow-up												
4	11027619	*rs4351014*	*HS3ST1*	C	T	0.684	0.94 [0.92-0.96]	$5.37x10^{-10}$	0.93 [0.88-0.98]	$4.54x10^{-3}$	0.94 [0.92-0.95]	$9.16x10^{-12}$
16	70694000	*rs4985556*	*IL34*	A	C	0.111	1.08 [1.05-1.11]	$2.28x10^{-8}$	1.09 [1.03-1.16]	$4.59x10^{-3}$	1.08 [1.06-1.11]	$3.91x10^{-10}$
16	81773209	*rs12444183*	*PLCG2*	A	G	0.407	0.95 [0.93-0.97]	$1.48x10^{-8}$	0.92 [0.88-0.96]	$3.23x10^{-5}$	0.95 [0.93-0.96]	$6.81x10^{-12}$

100-plus: 100-plus Study; *LASA*, Longitudinal aging study of Amsterdam; *ADC*, Amsterdam dementia cohort; *AGES*, Age/Gene Environment Susceptibility Study; *CEPH*, CEPH centenarian cohort; *CHS*, Cardiovascular Health Study; *DKLS*, Danish longevity study; *FHS*, Framingham Heart Study; *GEHA*, Genetics of Healthy Aging Study; *HRS*, Health and Retirement Study; *LLFS*, Long Life Family Study; *LLS*, Leiden Longevity Study; *Longevity*, Longevity Gene Project; *MrOS*, Osteoporotic Fractures in Men Study; *Newcastle 85+*, Newcastle 85+ Study; *RS*, Rotterdam study; *SOF*, Study of Osteoporotic Fracture; *Vitality 90+*, Vitality 90+ project; *GLS*, German longevity study; *CLHLS*, Chinese Longitudinal Healthy Longevity Survey. [a] For these studies, controls were provided by a separate cohort. Further details of the cohorts are provided in Supplementary Data 4.

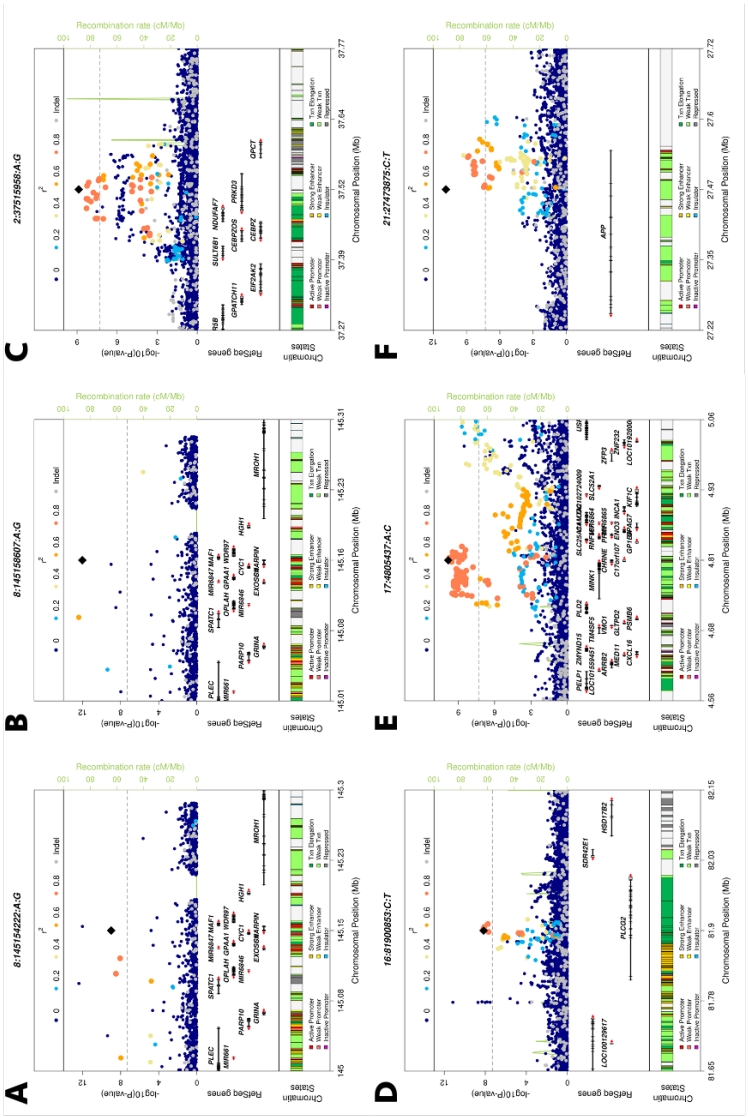


Figure 6.3: **Regional plots of the novel signals associated with AD in the overall meta-analysis. A-F.** The plots refer to the genomic regions surrounding *SHARPIN (1)*, *SHARPIN (2)*, *PRKD3*, *PLCG2*, *CHRNE* and *APP*, respectively.

we investigated the effect of sex and age at onset (AAO, Figure 6.5C). Our analysis revealed that the effect of the PRS was the same in both sexes (Figure 6.5C) and was consistent with both early-onset (onset before 65 years; *OR* = 1.58, *95% CI* [1.22–2.05], $p = 5.8\text{x}10^{-4}$) as well as with late-onset AD (onset later than 85 years; *OR* = 1.29, *95% CI* [1.10–1.51], $p = 1.5\text{x}10^{-3}$). PRSs has the potential to early identify subjects at risk of complex diseases.[23] To identify people at the highest genetic risk of AD based on the PRS, we used the validated 39-variants PRS in the large GR@ACE dataset. The PRS was associated with a 1.27-fold (*95% CI* [1.23–1.32]) increased risk for every standard deviation increase in the PRS ($p = 7.3\text{x}10^{-39}$) and with a gradual risk increase when we stratified the dataset into 2% percentiles of the PRS (Figure 6.6A, Supplementary Table 6). Next, we stratified the dataset in *APOE* genotype risk groups. The PRS percentiles were associated with AD within the *APOE* genotype groups (Figure 6.6B and Supplementary Table 7). Finally, we compared the risk extremes and found a 16.2-fold (*95% CI* [8.84–29.5], $p = 1.5\text{x}10^{-19}$) increased risk for the highest-PRS group (*APOE* $\epsilon 4\epsilon 4$) compared with the lowest-PRS group (*APOE* $\epsilon 2\epsilon 2/\epsilon 2\epsilon 3$; Supplementary Table 8). When we compared the median AAO in AD patients in these extreme risk groups we found a 9-year difference in the median age ($p_{Wilcoxon} = 1.7\text{x}10^{-6}$) (Figure 6.6C). Lastly, we studied the effects on AAO of the PRS in the *APOE* genotype groups. The PRS differentiated AAO only within *APOE* $\epsilon 4$ carriers. In *APOE* $\epsilon 4$ heterozygotes the PRS determined a 4-year difference in median AAO and in *APOE* $\epsilon 4$ homozygotes ($p_{Wilcoxon} = 6.9\text{x}10^{-5}$), where the PRS determined a median AAO difference of 5.5 years ($p_{Wilcoxon} = 4.6\text{x}10^{-5}$). For the selection of high-risk individuals, it is important to note that we found no difference in the odds and AAO for AD for *APOE* $\epsilon 4$ heterozygotes with the highest PRS compared to *APOE* $\epsilon 4$ homozygotes with the lowest PRS. The Cox regression also showed an impact of *APOE* on AAO, mainly on *APOE* $\epsilon 4\epsilon 4$ (significant *APOE*-PRS interaction ($p = 0.021$), Figure 6.6).

6.3 Discussion

This work adds on the ongoing global effort to identify genetic variants associated with AD (Figure 6.4). In the present work, we reported on the largest GWAS for AD risk to date, comprising genetic information of 467,623 individuals of European ancestry. We identified six novel variants that were not previously associated with the risk of AD and constructed a robust PRS for AD demonstrating its potential value for selecting subjects at risk of AD, especially within *APOE* $\epsilon 4$ carriers. This PRS was based on European ancestries and may or may not generalize to other ancestries. Validation in other populations will be required. We also acknowledge that controls included in GR@ACE are younger than cases and some of the controls might still develop AD later in life. This fact does not invalidate the analysis although reported estimates must be considered conservative. The differences in risk and AAO determined by the PRS of AD are relevant for design clinical trials that over-represent *APOE* $\epsilon 4$ carriers, as *APOE* $\epsilon 4$ heterozygous with highest PRS values have a similar risk and AAO to *APOE* $\epsilon 4$ homozygotes (Figure 6.6b). These represents $\sim$1% of our control population, which is the same percentage as all *APOE* $\epsilon 4$ homozygotes. A trial that aims to include *APOE* $\epsilon 4$ homozygotes, could consider widening the selection criteria and in this way hasten the enrollment process. Also, our PRS could aid at the interpretation of the results of clinical trials, as it determines a relevant proportion of the AAO, which could either mimic or obscure a treatment effect.

The most interesting finding from our GWAS is the discovery of a common protective (MAF (C-allele) = 0.483) intronic variant in the *APP* gene. Our results directly support *APP* production or processing as a causal pathway not only in familial AD but in common sporadic AD. The SNP is in a DNase hypersensitive area of 295 bp (chr21:27473781-27474075) possibly involved in the transcriptional regulation of the *APP* gene. *rs2154481* is an eQTL for the *APP* mRNA and an antisense transcript of the *APP* gene named AP001439.2 in public eQTL databases (Figure 6.10).[26] Functional evidence supports a modified *APP* transcription as an LD block of 13 SNPs within the *APP* locus (including *rs2154481*) increased the *TFCP2* transcription factor avidity to its binding site and increased the enhancer activity of this specific intronic region.[27] Based on this evidence, we can postulate that a life-long slightly higher *APP* gene expression protects the brain from AD insults. Still, this seems counterintuitive as duplications of the gene lead to early-onset AD.[28] A U-shaped effect, or hormesis effect of *APP* might help explain

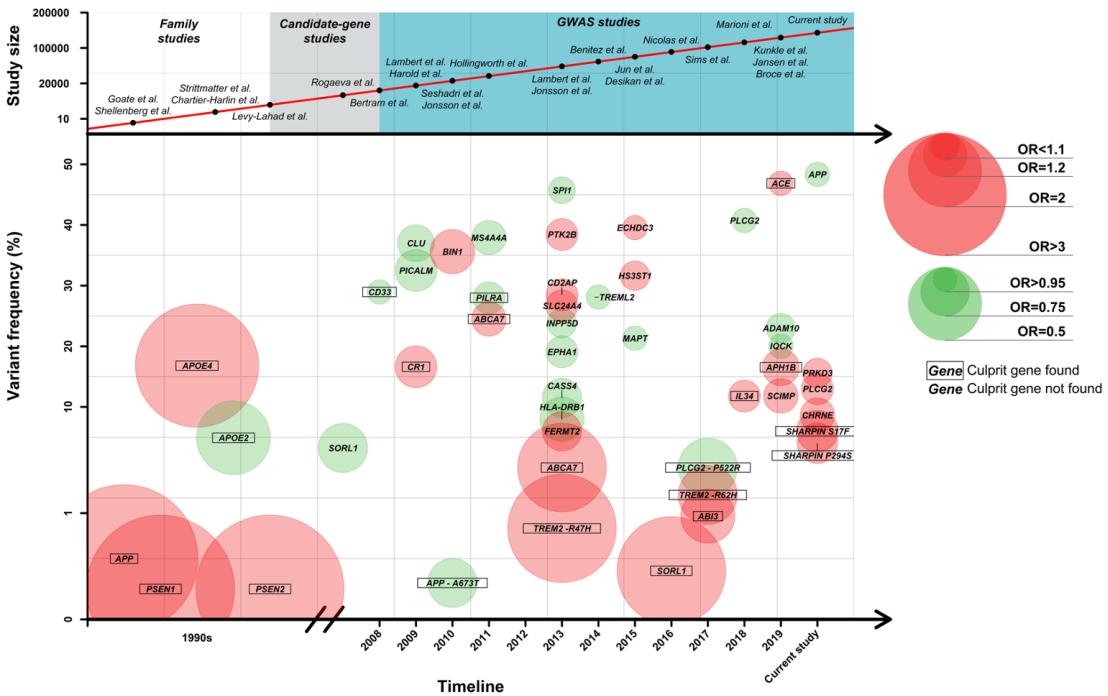


Figure 6.4: **Genetic landscape for Alzheimer's disease.** This figure shows the history of genetic discoveries in AD research over the past 30 years. This figure was constructed to our best knowledge of literature, but is not a systematic review of literature. For common variants, we selected only signals firmly replicated in large meta-GWAS (Lambert *et al.* 2013,[3] Kunkle *et al.* 2019,[19] Jun *et al.* 2016,[24] Sims *et al.* 2017,[6] Jansen *et al.* 2019,[25] and present study). For rare variants, we only selected those variants widely replicated excluding those loci presenting conflicting results. Abbreviations and more information about the genes can be found in Supplementary Table 4. *GWAS*, Genome-Wide Association Study; *OR*, Odds Ratio.

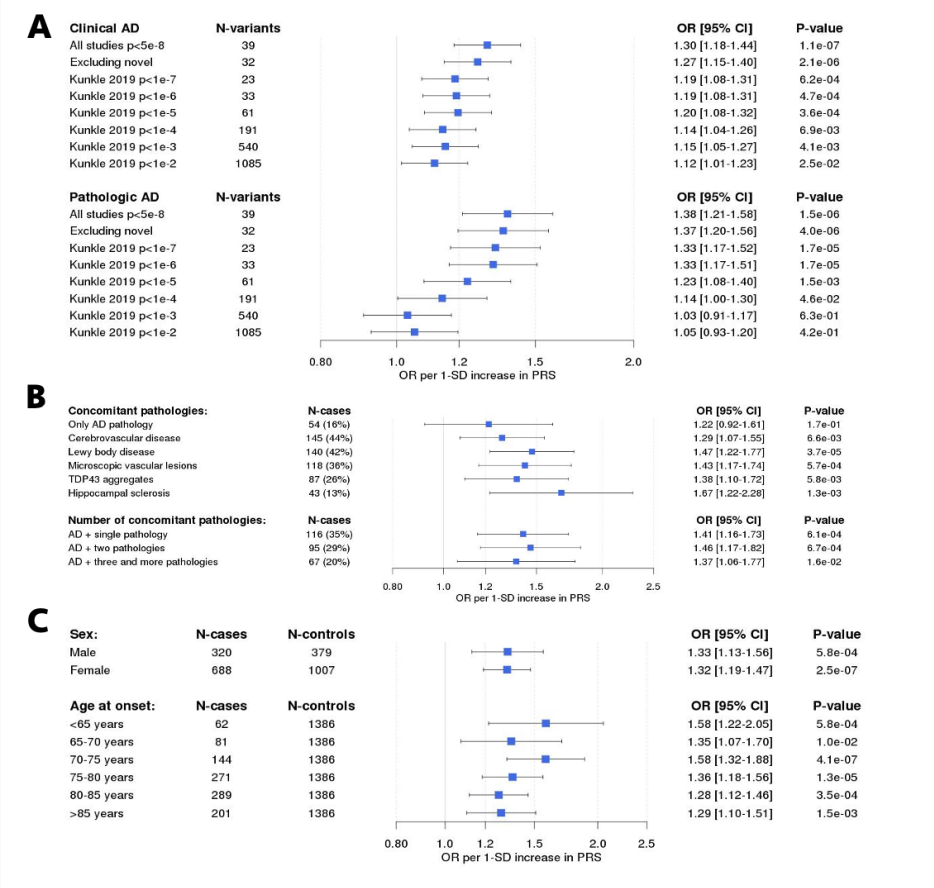


Figure 6.5: **Polygenic Risk Scores for AD. A.** The 39-SNP PRS association with clinical (OR = 1.38, per 1-SD increase in the PRS, 95% CI [1.21-1.58], $p = 1.5\text{x}10^{-6}$) and pathologically confirmed AD cases (OR = 1.30, 95% CI [1.18-1.44], $p = 1.1\text{x}10^{-7}$). **B.** PRS association with AD in the presence of concomitant brain pathologies (besides AD). **C.** PRS association with AD stratified by sex and AAO. A similar association of the PRS with AD was found in both sexes (OR_{males} = 1.33, [1.13-1.56], $p = 5.8\text{x}10^{-4}$ *vs.* $OR_{females}$ = 1.32, [1.19-1.47], $p = 2.5\text{x}10^{-7}$).

our observations and it might also fit the accelerated cognitive deterioration observed in AD patients treated with β-secretase inhibitors as these reduce β-amyloid in their brain.[29, 30] An alternative hypothesis is that mechanisms underlying the variant are related to the overexpression of protective fragments of the *APP* protein.[31] Disentangling the molecular mechanism

of our finding will help refine and steer the amyloid hypothesis.

Additionally, other three variants identified are altering protein sequence or affecting regulatory motifs. Two independent missense mutations in *SHARPIN* increased the AD risk. *SHARPIN* was previously proposed as an AD candidate gene, and functional analysis of a rare missense variant (NM_030974.3:p.Gly186Arg) resulted in the aberrant cellular localization of the variant protein and attenuated the activation of NF-κB, a central mediator of inflammatory and immune responses.[32, 33] Functional analysis of the two identified missense variants will show if the effect on immune reaction in AD is similar. The variant located in the *CHRNE* which encodes a subunit of the cholinergic receptor (*AChR*) is a strong modulator of *CHRNE* expression. The same allele that increases AD risk increases the expression in the brain and other tissues according to GTEx (p = 2.1x10^{-13}) (Figure 6.11). The detection of a potential hypermorph allele linked to AD risk and affecting cholinergic function could reintroduce this neurotransmitter pathway into the search for preventative strategies. Further functional studies are needed to consolidate this hypothesis. Altogether, we described six novel loci associated with sporadic AD. These signals reinforce that AD is a complex disease in which amyloid processing and immune response play key roles. We add to the growing body of evidence that the polygenic scores of all genetic loci to date, in combination with *APOE* genotypes, are robust tools that are associated with AD and its AAO. These properties make PRS promising in selecting individuals at risk to apply preventative therapeutic strategies.

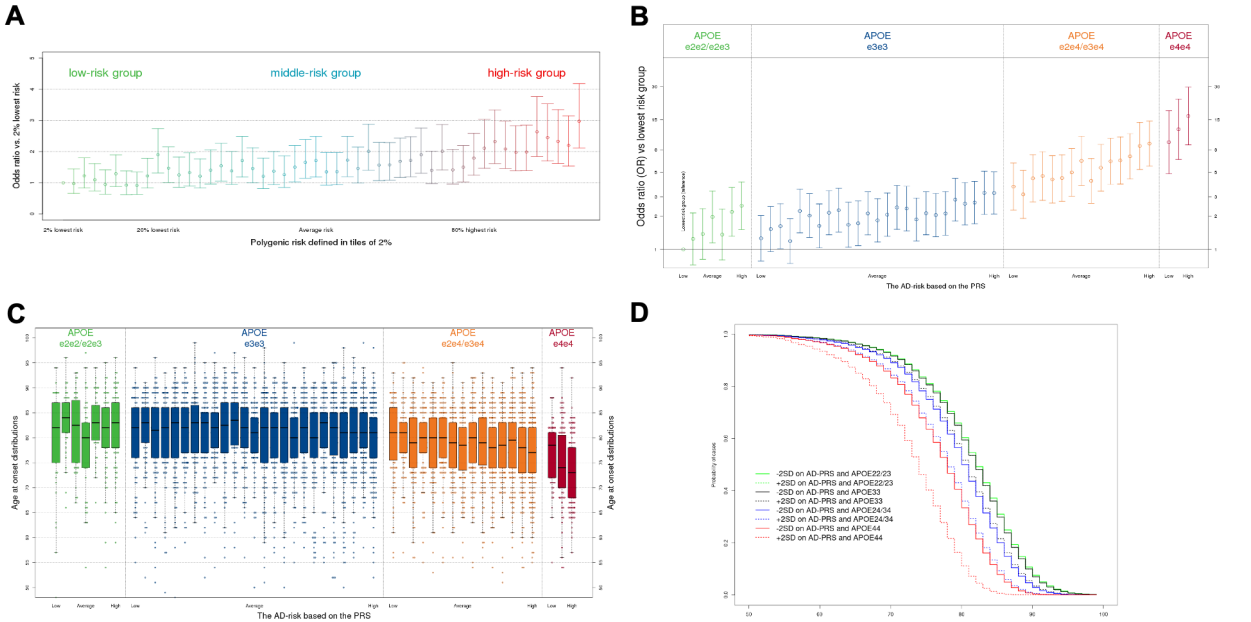


Figure 6.6: **Polygenic Risk Scores APOE stratification for AD. A.** The AD risk of PRS groups compared to those with the 2% lowest risk. There is a gradual risk increase over the percentiles. The 2% highest risk had a 3.0-fold (95% CI [2.12-4.18], p = 3.2x10^{-10}) increased risk compared with those with the 2% lowest risk. No interaction was found between the PRS and *APOE* genotypes (p-value=0.76). **B.** The AD risk stratified by PRS and *APOE* risk groups. The *APOE* $\epsilon 2\epsilon 2$ and $\epsilon 2\epsilon 3$ carriers as well as *APOE* $\epsilon 2\epsilon 4$ and $\epsilon 3\epsilon 4$ were pooled as they previously showed to have similar risk effects in AD. Association was found between highest and lowest PRS percentiles within the *APOE* genotype groups: $\epsilon 2\epsilon 2/\epsilon 2\epsilon 3$ carriers (OR = 2.48 [1.51-4.08], p = 3.4x10^{-4}), $\epsilon 3\epsilon 3$ carriers (OR = 2.67 [1.93-3.69], p = 3.5x10^{-9}), $\epsilon 2\epsilon 4/\epsilon 3\epsilon 4$ carriers (OR = 2.47 [1.67-3.66], p = 6.8x10^{-6}) and $\epsilon 4\epsilon 4$ carriers (OR = 2.02 [1.05-3.85], p = 3.4x10^{-2}). Comparisons of the highest and lowest PRS percentiles with respect to the *APOE* genotype groups: a difference was found between highest $\epsilon 2\epsilon 2/\epsilon 2\epsilon 3$ carriers *vs.* lowest $\epsilon 3\epsilon 3$ carriers (OR = 0.51 [0.34-0.75], p = 7.8x10^{-4}), but not between highest $\epsilon 3\epsilon 3$ carriers *vs.* lowest $\epsilon 2\epsilon 4/\epsilon 3\epsilon 4$ carriers (OR = 1.17 [0.82-1.66], p = 0.40) and highest $\epsilon 2\epsilon 4/\epsilon 3\epsilon 4$ carriers *vs.* lowest $\epsilon 4\epsilon 4$ carriers (OR = 0.89 [0.52-1.53], p = 0.68). **C.** The AAO of AD stratified by PRS and *APOE* risk groups. No difference in odds for AD was found between the PRS percentiles with AAO in *APOE* $\epsilon 2\epsilon 2/\epsilon 2\epsilon 3$ (lowest = 82 years, highest = 83 years, $p_{Wilcoxon}$ = 0.39) and *APOE* $\epsilon 3\epsilon 3$ (lowest = 82 years, highest = 81 years, p = 0.16). However, a 4-year difference was found between *APOE* $\epsilon 4$ heterozygotes ($p_{Wilcoxon}$ = 6.9x10^{-5}, 81 years compared with 77 years) and 5.5 years difference ($p_{Wilcoxon}$ = 4.6x10^{-5}, 78.5 years compared with 73 years) in *APOE* $\epsilon 4$ homozygotes. **D.** Cox regression model on AAO.

6.4 Methods

6.4.1 Samples and cohorts

Participants in this study were obtained from multiple sources, including raw data from case-control samples collected by GR@ACE/DEGESCO, summary statistics data from the case-control samples in the IGAP and the summary statistics of AD-by-proxy phenotype from the UK Biobank (see Supplementary Methods). An additional case-control samples from 16 independent cohorts (19,087 AD cases and 39,101 controls) was used for replication, largely collected and analyzed by the European Alzheimer's Disease Biobank (JPND-EADB) project. Full descriptions of the samples and their respective phenotyping and genotyping procedures are provided in the Supplementary Methods.

6.4.2 Meta-GWAS of AD

After study-specific variant filtering and quality-control procedures, we performed a fixed-effects inverse-variance-weighted meta-analysis on the discovery and follow-up stages.[21] To determine the lead SNPs (those with the strongest association per genomic region), we performed clumping on SNPs with a genome-wide significant p-value (p<5x10^{-8}) (PLINK v1.90, maximal linkage disequilibrium (LD) with R^2 <0.001 and physical distance of 250Kb). In the *APOE* region, we only considered the *APOE* $\epsilon 4$ (*rs429358*) and *APOE* $\epsilon 2$ (*rs7412*) SNPs.[34] LD information was calculated using the GR@ACE imputed genotypes as a reference. Polygenicity and confounding biases, such as cryptic relatedness and population stratification, can yield an inflated distribution of test statistics in GWAS. To distinguish between inflation from a true polygenic signal and bias we quantified the contribution of each by examining the relationship between test statistics and linkage disequilibrium (LD) using the LD Score regression intercept (LDSC software).[35] Chromosomal regions associated with AD in previous studies were excluded from follow-up.[3, 19, 25] We tested all variants with suggestive association (p<10^{-5}) located in proximity (200 Kb) of genomic regions selected for follow-up to allow for the potential refinement of the top associated variant. Conditional analyses were performed in regions where multiple variants were associated with AD using logistic regression models, adjusting for the genetic variants in the region. Regional plots were generated with a mixture of homemade Python (v2.7) and R (v3.6.0) scripts. Briefly, given an input variant, we calculated the LD between the input variant and all the surrounding variants within a window of length defined by the user. The LD

was calculated in the 1000Genomes samples of European ancestry. We used gene positions from RefSeq (release 93); in the case of multiple gene models for a given gene, we reported the model with the largest number of exons. We used recombination rates from HapMap II and chromatin states from ENCODE/Broad (15 states were grouped to highlight the predicted functional elements). As a reference genome, we used GRCh37. Quantile-quantile (QQ) plots, Manhattan plots, and the exploration of genomic inflation factors were performed using the R package *qqman*.

6.4.3 Polygenic Risk Score

We calculated a weighted individual PRS based on the 39 genetic variants that showed genome-wide significant (GWS) evidence of association with AD in the present study, excluding *APOE* to check the impact of PRS modulating *APOE* risk (Table 6.1 and Supplementary Table 3). The selected variants were directly genotyped or imputed with high quality (median imputation score R^2 = 0.93). The PRSs were generated by multiplying the genotype dosage of each risk allele for each variant by its respective weight and then summing across all variants. We weighted this by the effect size from previous IGAP studies: Kunkle *et al.* (36 variants),[19] Sims *et al.* (2 variants),[6] Jun *et al.* (*MAPT* locus),[24] Supplementary Table 5. The newly generated PRS was validated using logistic regression adjusted by four principal components in a sample of 676 AD cases diagnosed based on clinical criteria and 332 pathologically confirmed AD cases from the European Alzheimer's Disease Biobank-Fundació ACE/Barcelona Brain Bank dataset (EADB-F.ACE/BBB, Supplementary Information). This dataset was not used in prior genetic studies. In this dataset, all pathologically confirmed cases were scored for the presence or absence of concomitant pathologies. In all analyses, we compared the AD patients to the same control dataset (*N*=1,386). We performed analyses to test the robustness of the PRS. We tested the effect of adding variants below the genome-wide significance threshold using a pruning and thresholding approach. For this, we used the summary statistics of the IGAP study,[19] and we selected independent variants using the *clump_data()* function from the *TwoSampleMR* R-package (v0.4.25). We used strict settings for clumping (R^2 = 0.001 and window = 1 MB) and increasing *p*-value thresholds (>1x10^{-7}, >1x10^{-6}, >1x10^{-5}, >1x10^{-4}, >1x10^{-3}, >1x10^{-2}). We tested the association of the results with clinically diagnosed and pathologically confirmed AD patients. To evaluate the effect of diagnostic certainty, we tested whether the PRS was different between the two patient groups. For the PRS with 39

genome-wide significant variants, we tested whether the PRS had sex-specific effects, whether it resulted in different age-of-onset groups of AD, and the effect of the PRS in the presence of concomitant brain pathologies. Risk stratification of the validated PRSs. We searched for the groups at the highest risk of AD in the GR@ACE dataset (6,331 AD cases and 6,055 controls). We stratified the population into PRS percentiles, taking into account survival bias anticipated at old age.[17] To eliminate selection bias, we calculated the boundaries of the percentiles in the control participants aged 55 years and younger (*N*=3,546). Based on the boundaries from this population, the rest of the controls and all AD cases were then assigned into their appropriate percentiles. We first explored risk stratification using only the PRSs. For this, we split the PRSs into 50 groups (2 percentiles) and compared all groups with that which had the lowest PRS. Second, we explored risk stratification considering both the *APOE* genotypes and the PRSs. The *APOE* genotypes were pooled in the analyses as *APOE* $\epsilon 2\epsilon 2/\epsilon 2\epsilon 3$ (*N*=998, split into 7 PRS groups), *APOE* $\epsilon 3\epsilon 3$ (*N*=7,611, split into 25 PRS groups), *APOE* $\epsilon 2\epsilon 4/\epsilon 3\epsilon 4$ (*N*=3,399, split into 15 PRS groups) and *APOE* $\epsilon 4\epsilon 4$ (*N*=382, split into 3 PRS groups). We studied the effect of PRS across groups of individuals stratified by the *APOE* genotypes with the lowest PRS group (*APOE* as the reference group using logistic regression models adjusted for four population ancestry components). Finally, we compared the median age at onset using a Wilcoxon test. We implemented a Cox regression model on the GR@ACE/DEGESCO dataset case-only adjusted for covariates as *APOE* group, the interaction between the PRS and *APOE* and four population ancestry components. All analyses were done in R (v3.4.2).

6.4.4 Functional annotation

We used Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA, v1.3.4c) to interpret SNP-trait associations (see Supplementary Methods).[22] FUMA is an online platform that annotates GWAS findings and prioritizes the most likely causal SNPs and genes using information from 18 biological data repositories and tools. As input, we used the summary statistics of our meta-GWAS. Gene prioritization is based on a combination of positional mapping, expression quantitative trait loci (eQTL) mapping, and chromatin interaction mapping. Functional annotation was performed by applying a methodology similar to that described by Jansen *et al.*[25] We referred to the original publication for details on the methods and repositories of FUMA.[22]

6.4.5 Data availability

Summary statistics will be made available for download upon publication (`www.niagads.org`).

6.5 Acknowledgements

We would like to thank patients and controls who participated in this project. The present work has been performed as part of the doctoral program of I. de Rojas at the Universitat de Barcelona (Barcelona, Spain). The Genome Research @ Fundació ACE project (GR@ACE) is supported by Grifols SA, Fundación bancaria 'La Caixa', Fundació ACE, and CIBERNED. A.R. and M.B. receive support from the European Union/EFPIA Innovative Medicines Initiative Joint undertaking ADAPTED and MOPEAD projects (grant numbers 115975 and 115985, respectively). M.B. and A.R. are also supported by national grants PI13/02434, PI16/01861, PI17/01474 and PI19/01240. Acción Estratégica en Salud is integrated into the Spanish National R + D + I Plan and funded by ISCIII (Instituto de Salud Carlos III)–Subdirección General de Evaluación and the Fondo Europeo de Desarrollo Regional (FEDER–'Una manera de hacer Europa'). Some control samples and data from patients included in this study were provided in part by the National DNA Bank Carlos III (`www.bancoadn.org`, University of Salamanca, Spain) and Hospital Universitario Virgen de Valme (Sevilla, Spain); they were processed following standard operating procedures with the appropriate approval of the Ethical and Scientific Committee.

Amsterdam dementia Cohort (ADC): Research of the Alzheimer center Amsterdam is part of the neurodegeneration research program of Amsterdam Neuroscience. The Alzheimer Center Amsterdam is supported by Stichting Alzheimer Nederland and Stichting VUmc fonds. The clinical database structure was developed with funding from Stichting Dioraphte. Genotyping of the Dutch case-control samples was performed in the context of EADB (European Alzheimer DNA biobank) funded by the JPco-fuND FP-829-029 (ZonMW project number 733051061). 100-Plus study: We are grateful for the collaborative efforts of all participating centenarians and their family members and/or relations. This work was supported by Stichting Alzheimer Nederland (WE09.2014-03), Stichting Diorapthe, horstingstuit foundation, Memorabel (ZonMW project number 733050814, 733050512) and Stichting VUmc Fonds. Genotyping of the 100-Plus Study was performed in the context of EADB (European Alzheimer DNA biobank) funded by the JPcofuND FP-829-029 (ZonMW project number 733051061). Longitudinal Aging Study

Amsterdam (LASA) is largely supported by a grant from the Netherlands Ministry of Health, Welfare and Sports, Directorate of Long-Term Care. The authors are grateful to all LASA participants, the fieldwork team and all researchers for their ongoing commitment to the study.

This work was supported by a grant (European Alzheimer DNA BioBank, EADB) from the EU Joint Programme – Neurodegenerative Disease Research (JPND) and also funded by Inserm, Institut Pasteur de Lille, the Lille Métropole Communauté Urbaine, the French government's LABEX DISTALZ program (development of innovative strategies for a transdisciplinary approach to Alzheimer's disease). Genotyping of the German case-control samples was performed in the context of EADB (European Alzheimer DNA biobank) funded by the JPcofuND (German Federal Ministry of Education and Research, BMBF: 01ED1619A). Full acknowledgments for the studies that contributed data can be found in the Supplementary Note. We thank the numerous participants, researchers, and staff from many studies who collected and contributed to the data.

We thank the International Genomics of Alzheimer's Project (IGAP) for providing summary results data for these analyses. The investigators within IGAP contributed to the design and implementation of IGAP and/or provided data but did not participate in analysis or writing of this report. IGAP was made possible by the generous participation of the control subjects, the patients, and their families. The i–Select chips was funded by the French National Foundation on Alzheimer's disease and related disorders. EADI was supported by the LABEX (laboratory of excellence program investment for the future) DISTALZ grant, Inserm, Institut Pasteur de Lille, Université de Lille 2 and the Lille University Hospital. GERAD was supported by the Medical Research Council (Grant n° 503480), Alzheimer's Research UK (Grant n° 503176), the Wellcome Trust (Grant n° 082604/2/07/Z) and German Federal Ministry of Education and Research (BMBF): Competence Network Dementia (CND) grant n° 01GI0102, 01GI0711, 01GI0420. CHARGE was partly supported by the NIH/NIA grant R01 AG033193 and the NIA AG081220 and AGES contract N01–AG–12100, the NHLBI grant R01 HL105756, the Icelandic Heart Association, and the Erasmus Medical Center and Erasmus University. ADGC was supported by the NIH/NIA grants: U01 AG032984, U24 AG021886, U01 AG016976, and the Alzheimer's Association grant ADGC–10–196728.

This research has been conducted using the UK Biobank public resource obtained through the University of Edinburg Data Share (`https://datashare.is.ed.ac.uk/handle/10283/3364`).

6.6 Full author list and affiliations

Itziar de Rojas,[1,2] Sonia Moreno-Grau,[1,2] Niccolo' Tesi,[3,4] Benjamin Grenier-Boley,[5] Victor Andrade,[6,7] Iris Jansen,[3,8] Nancy L. Pedersen,[9] Najada Stringa,[10] Anna Zettergren,[11] Isabel Hernàndez,[1,2] Laura Montrreal,[1] Carmen Antùnez,[12] Anna Antonell,[13] Rick M. Tankard,[14] Joshua C. Bis,[15] Rebecca Sims,[16,17] Cèline Bellenguez,[5] Inès Quintela,[18] Antonio Gonzàlez-Perez,[19] Miguel Calero,[20,21,2] Emilio Franco,[22] Juan Macìas,[23] Rafael Blesa,[24,2] Manuel Menéndez-Gonzàlez,[25,26] Ana Frank-Garcìa,[27,28,29,2] Jose Luìs Royo,[30] Fermìn Moreno,[31,2] Raquel Huerto,[32,33] Miquel Baquero,[34] Mònica Diez-Fairen,[35] Carmen Lage,[36,2] Sebastian Garcia-Madrona,[37] Pablo Garcìa,[1] Emilio Alarcòn-Martìn,[30,1] Sergi Valero,[1,2] Oscar Sotolongo-Grau,[1] EADB, GR@ACE, DEGESCO, IGAP (ADGC, CHARGE, EADI, GERAD) and PGC-ALZ Consortia, Guillermo Garcia-Ribas,[37] Pascual Sànchez-Juan,[36,2] Pau Pastor,[35] Jordi Pèrez-Tur,[38,34,2] Gerard Pinol-Ripoll,[32,33] Adolfo Lopez de Munain,[31,39] Jose Marìa Garcìa-Alberca,[40] Marìa J. Bullido,[41,28,29,2] Victoria Alvarez,[25,26] Alberto Lleò,[24,2] Luis M. Real,[23,42] Pablo Mir,[43,2] Miguel Medina,[2,21] Philip Scheltens,[10] Henne Holstege,[10,4] Marta Marquiè,[1] Marìa Eugenia Sàez,[19] Angel Carracedo,[18,44] Philippe Amouyel,[5] Julie Williams,[16,17] Sudha Seshadri,[45,46,47] Cornelia M. van Duijn,[48] Karen A. Mather,[49,50] Raquel Sànchez-Valle,[13] Manuel Serrano-Rìos,[51] Adelina Orellana,[1,2] Lluìs Tàrraga,[1,2] Kaj Blennow,[52,53] Martijn Huisman,[10,54] Ole A. Andreassen,[55,56] Danielle Posthuma,[8,57] Jordi Clarimòn,[24,2] Mercè Boada,[1,2] Wiesje M. van der Flier,[3] Alfredo Ramirez,[6,7,58] Jean-Charles Lambert,[5] Sven J. van der Lee,[3,4] and Agustìn Ruiz [1,2]

[1] Research Center and Memory clinic Fundaciò ACE, Institut Català de Neurociències Aplicades, Universitat Internacional de Catalunya, Barcelona, Spain.

[2] CIBERNED, Network Center for Biomedical Research in Neurodegenerative Diseases, National Institute of Health Carlos III, Madrid, Spain.

[3] Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands.

[4] Department of Clinical Genetics, VU University Medical Centre, Amsterdam, The Netherlands.

[5] Univ. Lille, Inserm, Institut Pasteur de Lille, CHU Lille, U1167 - Labex DISTALZ - RID-AGE - Risk factors and molecular determinants of aging-related diseases, F-59000 Lille, France.

[6] Division of Neurogenetics and Molecular Psychiatry, Department of Psychiatry and Psychotherapy, University of Cologne, Medical Faculty, 50937 Cologne, Germany.

[7] Department of Neurodegeneration

and Geriatric Psychiatry, University of Bonn, 53127 Bonn, Germany.
[8] Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands.
[9] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.
[10] Amsterdam UMC - Vrije Universiteit Amsterdam, Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands.
[11] Neuropsychiatric Epidemiology Unit, Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, Sahlgrenska Academy, Centre for Ageing and Health (AgeCap) at the University of Gothenburg,Sweden.
[12] Unidad de Demencias, Hospital Clìnico Universitario Virgen de la Arrixaca, Murcia, Spain.
[13] Alzheimers disease and other cognitive disorders unit. Service of Neurology. Hospital Clìnic of Barcelona. Institut d'Investigacions Biomèdiques August Pi i Sunyer, University of Barcelona, Barcelona, Spain.
[14] Mathematics and Statistics, Murdoch University, WA, Australia.
[15] Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA.
[16] Division of Psychological Medicine and Clinial Neurosciences, MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, UK.
[17] UK Dementia Research Institute at Cardiff, Cardiff University, Cardiff, UK.
[18] Grupo de Medicina Xenòmica, Centro Nacional de Genotipado (CEGEN-PRB3-ISCIII). Universidad de Santiago de Compostela, Santiago de Compostela, Spain.
[19] CAEBI, Centro Andaluz de Estudios Bioinformàticos, Sevilla, Spain.
[20] UFIEC, Instituto de Salud Carlos III.
[21] CIEN Foundation/Queen Sofia Foundation Alzheimer Center.
[22] Unidad de Demencias, Servicio de Neurologìa y Neurofisiologìa. Instituto de Biomedicina de Sevilla (IBiS), Hospital Universitario Virgen del Rocìo/CSIC/Universidad de Sevilla, Seville, Spain.
[23] Unidad Clìnica de Enfermedades Infecciosas y Microbiologìa. Hospital Universitario de Valme, Sevilla, Spain.
[24] Department of Neurology, II B Sant Pau, Hospital de la Santa Creu i Sant Pau, Universitat Autònoma de Barcelona, Barcelona, Spain.
[25] Hospital Universitario Central de Asturias, Oviedo, Spain.
[26] Instituto de Investigaciòn Sanitari del Principado de Asturias.
[27] Hospital Universitario la Paz.
[28] Instituto de Investigaciòn Sanitaria Hospital la Paz (IdIPaz), Madrid, Spain.
[29] Universidad Autònoma de Madrid.
[30] Department of Surgery, Biochemistry and Molecular Biology, School of Medicine, University of Màlaga, Màlaga, Spain.
[31] Department of Neurology. Hospital

Universitario Donostia. San Sebastian, Spain.
[32] Unitat Trastorns Cognitius, Hospital Universitari Santa Maria de Lleida, Lleida, Spain.
[33] Institut de Recerca Biomèdica de Lleida (IRBLLeida), Lleida, Spain.
[34] Servei de Neurologia, Hospital Universitari i Politècnic La Fe, Valencia, Spain.
[35] Fundaciò Docència i Recerca MùtuaTerrassa and Movement Disorders Unit, Department of Neurology, University Hospital Mùtua Terrassa, Terrassa 08221, Barcelona, Spain.
[36] Neurology Service, Marquès de Valdecilla University Hospital (University of Cantabria and IDIVAL), Santander, Spain.
[37] Hospital Universitario Ramon y Cajal, IRYCIS, Madrid.
[38] Unitat de Genètica Molecular, Institut de Biomedicina de València-CSIC, Valencia, Spain.
[39] Department of Neurosciences. Faculty of Medicine and Nursery. University of the Basque Country, San Sebastiàn, Spain.
[40] Alzheimer Research Center and Memory Clinic, Andalusian Institute for Neuroscience, Màlaga, Spain.
[41] Centro de Biologìa Molecular Severo Ochoa (UAM-CSIC).
[42] Depatamento de Especialidades Quirùrgicas, Bioquìmica, Biologìa Molecular e Inmunologìa. Facultad de Medicina. Universidad de Màlaga. Màlaga (Spain).
[43] Unidad de Trastornos del Movimiento, Servicio de Neurologìa y Neurofisiologìa. Instituto de Biomedicina de Sevilla (IBiS), Hospital Universitario Virgen del Rocìo/CSIC/Universidad de Sevilla, Seville, Spain.
[44] Fundaciòn Pùblica Galega de Medicina Xenòmica- CIBERER-IDIS, Santiago de Compostela, Spain.
[45] Glenn Biggs Institute for Alzheimer's and Neurodegenerative Diseases, San Antonio, TX, USA.
[46] Framingham Heart Study, Framingham, MA, USA.
[47] Department of Neurology, Boston University School of Medicine, Boston, MA, USA.
[48] Nuffield Department of Population Health Fellow of St Cross college.
[49] Centre for Healthy Brain Ageing (CHeBA), School of Psychiatry, Faculty of Medicine, University of New South Wales, Sydney 2052, Australia.
[50] Neuroscience Research Australia, Sydney, NSW, Australia.
[51] Centro de Investigaciòn Biomèdica en Red de Diabetes y Enfermedades Metabòlicas Asociadas, CIBERDEM, Spain, Hospital Clìnico San Carlos, Madrid, Spain.
[52] Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital, Molndal, Sweden.
[53] Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology, Sahlgrenska Academy at the University of Gothenburg, Sweden.
[54] Department of Sociology, VU University, Amsterdam, the Netherlands.
[55] NORMENT, K.G. Jebsen Centre for

Psychosis Research, Institute of Clinical Medicine, University of Oslo, Oslo, Norway.
[56] Institute of Clinical Medicine, University of Oslo, Oslo, Norway.
[57] University, Amsterdam, the Netherlands.
[58] German Center for Neurodegenerative Diseases (DZNE), 53127 Bonn, Germany.

6.7 Supplementary Figures

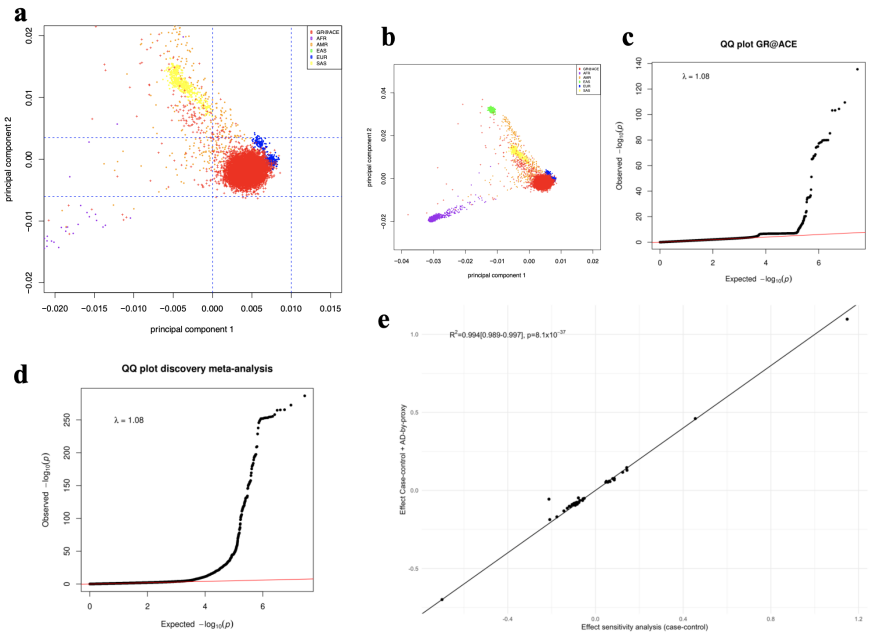


Figure 6.7: **Genome-wide association study. a-c**, Principal component analysis and QQ-plot for the GR@ACE dataset. **d**, QQplot Discovery meta-analysis.**e**, Correlation between the effect estimates from the AD case-control and AD by proxy approach for the significant loci. We compared the results obtained to a second meta-analysis using only the case-control datasets (IGAP Stages 1–2) and GR@ACE datasets as a sensitivity analysis to identify false negative results given possible dilution by the by- proxy approach in the UK Biobank (Supplementary Data 3). The meta-analysis, including the by-proxy summary statistics, identified 11 additional loci reaching genome-wide significance with respect to case-control-only results. The incorporation of by-proxy summary statistics did not show an association in two previously reported AD loci (*rs7185636-IQCK* and *rs386572859-MAPT*) by the IGAP consortium and replicated in the GR@ACE dataset (OR = 0.93[0.90-0.95], p = 4.5x10^{-8} and OR = 0.81[0.75-0.87], p = 7.9x10^{-9}, respectively).

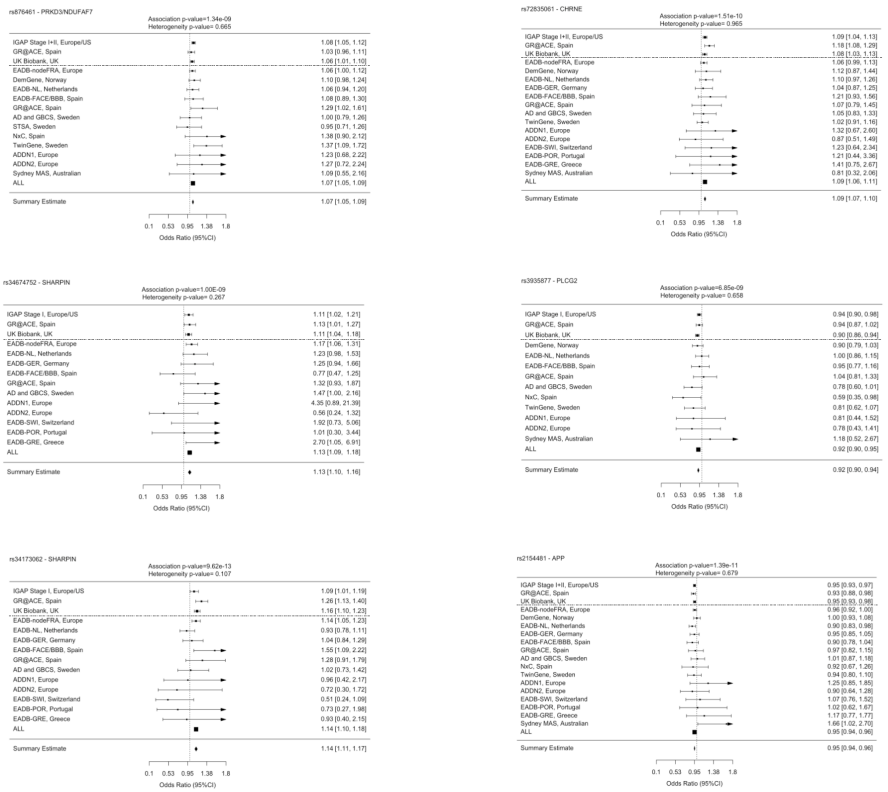


Figure 6.8: **Forest plots for the six novel signals identified in overall meta-analysis.** See sample size in Supplementary Data 1. Data are presented as Odds Ratio (95% CI).

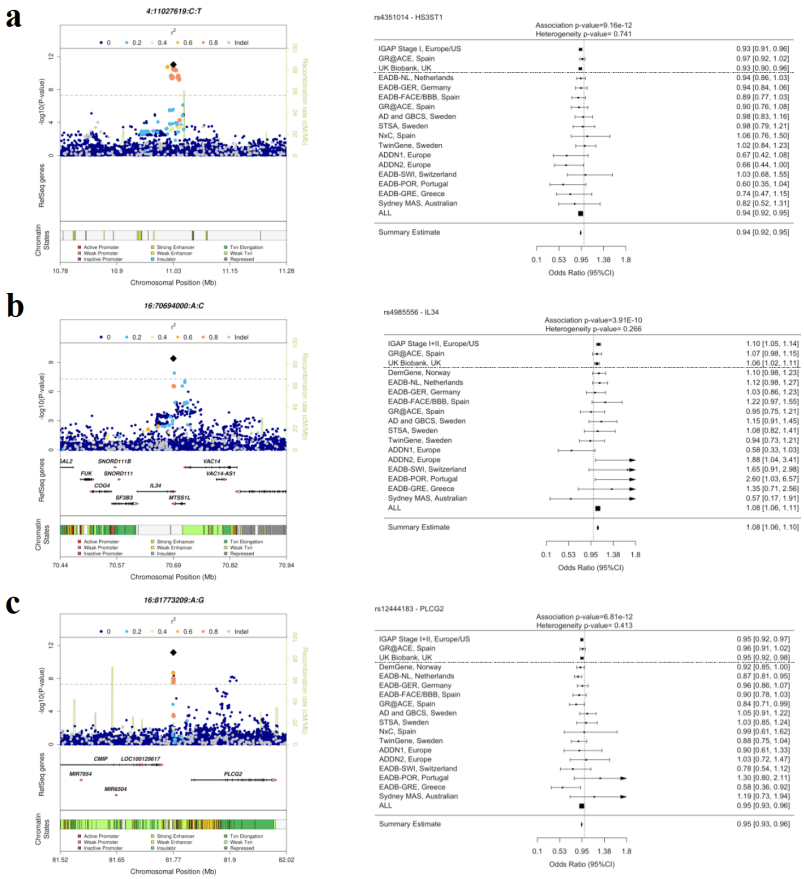


Figure 6.9: **LocusZoom and forest plots: strengthened evidence of association with AD for three additional genomic loci. a**, *HS3ST1* loci. **b**, *IL34* loci and **c**, *PLCG2* loci. Data for the forest plots are presented as Odds Ratio (95% CI). The first was *rs4351014* with AD (combined OR = 0.94 [0.92-0.95], p = 9.2x10^{-12}). This variant has been previously linked to *HS3ST1*. The second was a stop-codon mutation (*rs4985556*, Tyr213Ter, MAF = 0.111) in the interleukin 34 (*IL34*) gene that was previously reported in a by-proxy approach (combined OR = 1.08 [1.06-1.11], p = 3.9x10^{-10}). The third genomic region contains the *PLCG2* gene, which has been associated with AD twice before (the rare missense variant p.P522R in the *PLCG2* gene and *rs12444183* near the promotor region of *PLCG2*). After the combination of discovery and follow-up, a third independent association signal emerged in the *PLCG2* region (*rs3935877*, effect allele frequency = 0.868, OR = 0.92 [0.90-0.95], p = 6.9x10^{-9}). We also strengthened the association of *PLCG2-rs12444183* with AD (MAF = 0.407, combined OR = 0.95 [0.93-0.96], p = 6.8x10^{-12}). Conditional analyses in the *PLCG2* region showed that the association signals of all three variants are independent. A conditional analysis on the nearby *SCIMP* locus (333 Kb) (Supplementary Data 14) showed similar effects after adjustment for *SCIMP*, in line with the fact that the two independent signals are in weak LD (R^2 = 0.139, D' = 0.446).

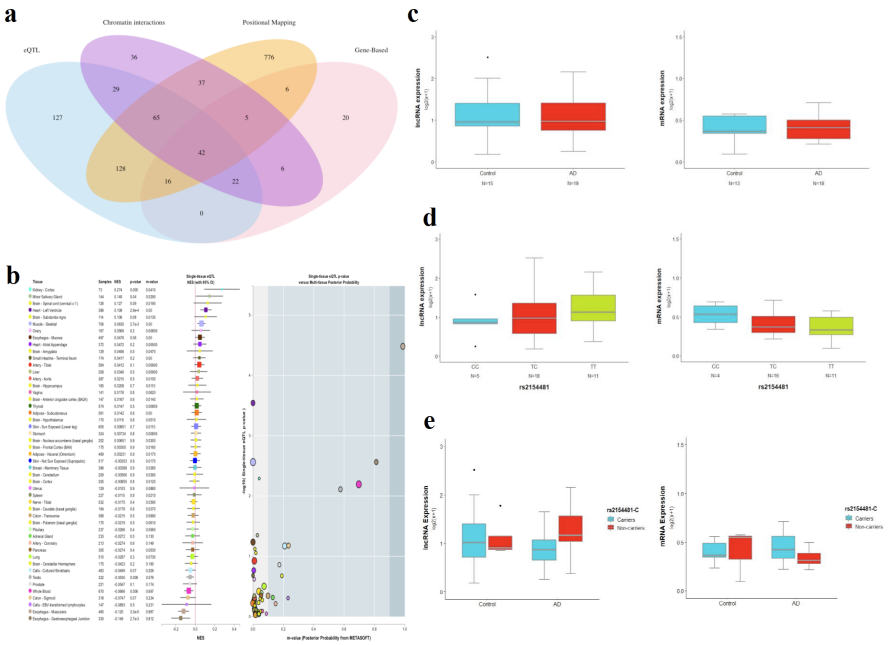


Figure 6.10: **Functional analysis.** **a** Diagram of functional interpretation by 4 FUMA strategies. To link the novel variants to specific genes and functional motifs in their genomic regions, we applied different strategies implemented on the FUMA platform. FUMA helps to generate hypotheses that are testable in functional experiments aimed at proving causal relations. The genes *APP, IL34, CHRNE, PLCG2,* and *SHARPIN* were the most likely candidate genes in the regions as they were implicated in at least three mapping strategies (Supplementary Data 15-18). **b**, Differential tissue expression for *APP* eQTL according to GTEx. Data are presented as Normalized effect size (NES) with 95% CI. **c**, Differential expression of lncAPP and mRNA in AD cases and controls (n=34 lncAPP and n=31 mRNA biologically independent samples). No significant differences were found between the total expression of the lncAPP (AP001439.2) and mRNA expression in the brain case/control samples. **d**, Differential expression of lncAPP and mRNA stratified by genotype. The allelic frequency was as expected (MAF = 0.41), as well as the eQTL effect for mRNA (CC>TC>TT) and lncAPP (CC<TC<TT) according to GTEx. **e**, Expression of lncAPP and mRNA stratified by *rs2154481* allele C carriers or non-carriers in AD cases and controls respectively. Interestingly, we saw an increase in the expression of the lncAPP associated with the T allele that seems more exacerbated in the patients than in the controls. If so, the protective C allele (*rs2154481*) would also be associated with a decrease in the expression of the lncAPP, thus being able to modify the final expression of *APP*. In **c-e**, data are represented as boxplots where the middle line is the median, the lower and upper hinges correspond to the first and third quartiles, the whiskers extend to the hinge to the inter-quartile range (IQR), while data beyond the end of the whiskers are outlying points that are plotted individually according to the manual of *ggplot2* package in R.

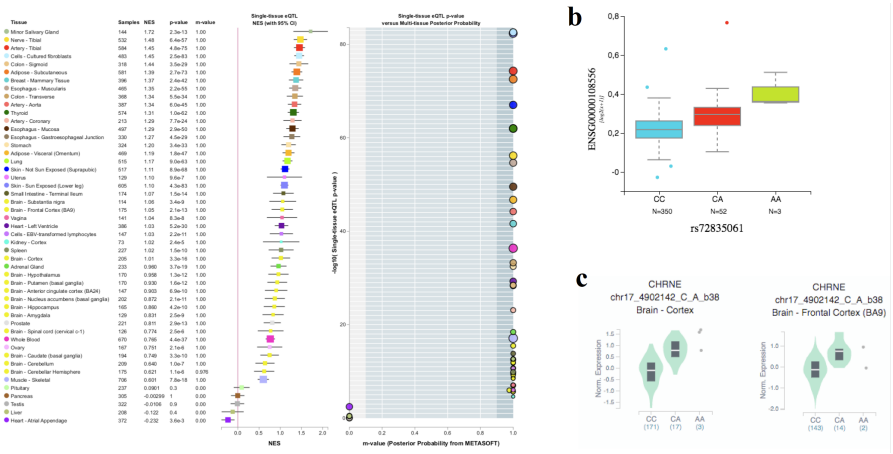


Figure 6.11: **Expression for CHRNE eQTL. a**, Differential tissue expression for *CHRNE* eQTL according to GTEX. Data are presented as Normalized effect size (NES) with 95% CI. **b-c**, Expression of the *CHRNE* transcript in the brain according to BrainSeq and GTEx respectively. Data are represented as boxplots where the middle line is the median, the lower and upper hinges correspond to the first and third quartiles, the whiskers extend to the hinge to the inter-quartile range (IQR), while data beyond the end of the whiskers are outlying points that are plotted individually according to the manual of *ggplot2* package in R

6.8 Supplementary Tables

Supplementary Tables and Supplementary Information can be accessed by scanning the following code or accessing the journal's website here.

References

[1] M. A. Pericak-Vance et al. "Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage". In: *American Journal of Human Genetics* 48.6 (June 1991), pp. 1034–1050. ISSN: 0002-9297.

[2] Sudha Seshadri et al. "Genome-wide analysis of genetic loci associated with Alzheimer disease". In: *JAMA* 303.18 (May 2010), pp. 1832–1840. ISSN: 1538-3598. DOI: 10.1001/jama.2010.574.

[3] J. C. Lambert et al. "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease". In: *Nature Genetics* 45.12 (Dec. 2013), pp. 1452–1458. ISSN: 1546-1718. DOI: 10.1038/ng.2802.

[4] Thorlakur Jonsson et al. "Variant of TREM2 associated with the risk of Alzheimer's disease". In: *The New England Journal of Medicine* 368.2 (Jan. 2013), pp. 107–116. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1211103.

[5] Rita Guerreiro et al. "TREM2 variants in Alzheimer's disease". In: *The New England Journal of Medicine* 368.2 (Jan. 2013), pp. 117–127. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1211851.

[6] Rebecca Sims et al. "Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease". In: *Nature Genetics* 49.9 (Sept. 2017), pp. 1373–1384. ISSN: 1546-1718. DOI: 10.1038/ng.3916.

[7] Perry G. Ridge et al. "Assessment of the genetic variance of late-onset Alzheimer's disease". In: *Neurobiology of Aging* 41 (May 2016), 200.e13–200.e20. ISSN: 1558-1497. DOI: 10.1016/j.neurobiolaging.2016.02.024.

[8] "2012 Alzheimer's disease facts and figures". In: *Alzheimer's & Dementia* 8.2 (Mar. 2012), pp. 131–168. ISSN: 15525260. DOI: 10.1016/j.jalz.2012.02.001.

[9] Eric M. Reiman et al. "Alzheimer's Prevention Initiative: a plan to accelerate the evaluation of presymptomatic treatments". In: *Journal of Alzheimer's disease: JAD* 26 Suppl 3 (2011), pp. 321–329. ISSN: 1875-8908. DOI: 10.3233/JAD-2011-0059.

[10] Li-Kai Huang, Shu-Ping Chao, and Chaur-Jong Hu. "Clinical trials of new drugs for Alzheimer disease". In: *Journal of Biomedical Science* 27.1 (Jan. 6, 2020), p. 18. ISSN: 1423-0127. DOI: 10.1186/s12929-019-0609-7.

[11] E. Rodríguez-Rodríguez et al. "Genetic risk score predicting accelerated progression from mild cognitive impairment to Alzheimer's disease". In: *Journal of Neural Transmission (Vienna, Austria: 1996)* 120.5 (May 2013), pp. 807–812. ISSN: 1435-1463. DOI: 10.1007/s00702-012-0920-x.

[12] Hieab H. H. Adams et al. "Genetic risk of neurodegenerative diseases is associated with mild cognitive impairment and conversion to dementia". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 11.11 (Nov. 2015), pp. 1277–1285. ISSN: 1552-5279. DOI: 10.1016/j.jalz.2014.12.008.

[13] Rahul S. Desikan et al. "Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score". In: *PLOS Medicine* 14.3 (Mar. 2017). Ed. by Carol Brayne, e1002258. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002258.

[14] A. Lacour et al. "Genome-wide significant risk factors for Alzheimer's disease: role in progression to dementia due to Alzheimer's disease among subjects with mild cognitive impairment". In: *Molecular Psychiatry* 22.1 (2017), pp. 153–160. ISSN: 1476-5578. DOI: 10.1038/mp.2016.18.

[15] Kristel Sleegers et al. "A 22-single nucleotide polymorphism Alzheimer's disease risk score correlates with family history, onset age, and cerebrospinal fluid Aβ42". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 11.12 (Dec. 2015), pp. 1452–1460. ISSN: 1552-5279. DOI: 10.1016/j.jalz.2015.02.013.

[16] Australian Imaging Biomarkers and Lifestyle (AIBL) Study et al. "Risk prediction of late-onset Alzheimer's disease implies an oligogenic architecture". In: *Nature Communications* 11.1 (Dec. 2020), p. 4799. ISSN: 2041-1723. DOI: 10.1038/s41467-020-18534-1.

[17] Sven J van der Lee et al. "The effect of APOE and other common genetic variants on the onset of Alzheimer's disease and dementia: a community-based cohort study". In: *The Lancet Neurology* 17.5 (May 2018), pp. 434–444. ISSN: 14744422. DOI: 10.1016/S1474-4422(18)30053-X.

[18] Sonia Moreno-Grau et al. "Genome-wide association analysis of dementia and its clinical endophenotypes reveal novel loci associated with Alzheimer's disease and three causality networks: The GR@ACE project". In: *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 15.10 (2019), pp. 1333–1347. ISSN: 1552-5279. DOI: 10.1016/j.jalz.2019.06.4950.

[19] Alzheimer Disease Genetics Consortium (ADGC), et al. "Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing". In: *Nature Genetics* 51.3 (Mar. 2019), pp. 414–430. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-019-0358-2.

[20] Riccardo E. Marioni et al. "GWAS on family history of Alzheimer's disease". In: *Translational Psychiatry* 8.1 (Dec. 2018), p. 99. ISSN: 2158-3188. DOI: 10.1038/s41398-018-0150-6.

[21] Cristen J. Willer, Yun Li, and Gonçalo R. Abecasis. "METAL: fast and efficient meta-analysis of genomewide association scans". In: *Bioinformatics (Oxford, England)* 26.17 (Sept. 1, 2010), pp. 2190–2191. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq340.

[22] Kyoko Watanabe et al. "Functional mapping and annotation of genetic associations with FUMA". In: *Nature Communications* 8.1 (Dec. 2017), p. 1826. ISSN: 2041-1723. DOI: 10.1038/s41467-017-01261-5.

[23] Amit V. Khera et al. "Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations". In: *Nature Genetics* 50.9 (2018), pp. 1219–1224. ISSN: 1546-1718. DOI: 10.1038/s41588-018-0183-z.

[24] G. Jun et al. "A novel Alzheimer disease locus located near the gene encoding tau protein". In: *Molecular Psychiatry* 21.1 (Jan. 2016), pp. 108–117. ISSN: 1476-5578. DOI: 10.1038/mp.2015.23.

[25] Iris E. Jansen et al. "Genome-wide meta-analysis identifies new loci and functional pathways influencing

Alzheimer's disease risk". In: *Nature Genetics* 51.3 (Mar. 2019), pp. 404–413. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-018-0311-9.

[26] GTEx Consortium. "The Genotype-Tissue Expression (GTEx) project". In: *Nature Genetics* 45.6 (June 2013), pp. 580–585. ISSN: 1546-1718. DOI: 10.1038/ng.2653.

[27] Craig Myrum et al. "Implication of the APP Gene in Intellectual Abilities". In: *Journal of Alzheimer's disease: JAD* 59.2 (2017), pp. 723–735. ISSN: 1875-8908. DOI: 10.3233/JAD-170049.

[28] Giovanna Cenini et al. "Wild type but not mutant APP is involved in protective adaptive responses against oxidants". In: *Amino Acids* 39.1 (June 2010), pp. 271–283. ISSN: 1438-2199. DOI: 10.1007/s00726-009-0438-1.

[29] Michael F. Egan et al. "Randomized Trial of Verubecestat for Mild-to-Moderate Alzheimer's Disease". In: *The New England Journal of Medicine* 378.18 (2018), pp. 1691–1703. ISSN: 1533-4406. DOI: 10.1056/NEJMoa1706441.

[30] Francesco Panza et al. "Do BACE inhibitor failures in Alzheimer patients challenge the amyloid hypothesis of the disease?" In: *Expert Review of Neurotherapeutics* 19.7 (2019), pp. 599–602. ISSN: 1744-8360. DOI: 10.1080/14737175.2019.1621751.

[31] Dimitri Hefter and Andreas Draguhn. "APP as a Protective Factor in Acute Neuronal Insults". In: *Frontiers in Molecular Neuroscience* 10 (2017), p. 22. ISSN: 1662-5099. DOI: 10.3389/fnmol.2017.00022.

[32] Daniel Lancour et al. "One for all and all for One: Improving replication of genetic studies through network diffusion". In: *PLoS genetics* 14.4 (2018), e1007306. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1007306.

[33] Yuya Asanomi et al. "A rare functional variant of SHARPIN attenuates the inflammatory response and associates with increased risk of late-onset Alzheimer's disease". In: *Molecular Medicine (Cambridge, Mass.)* 25.1 (2019), p. 20. ISSN: 1528-3658. DOI: 10.1186/s10020-019-0090-5.

[34] Gyungah Jun et al. "Comprehensive search for Alzheimer disease susceptibility loci in the APOE region". In: *Archives of Neurology* 69.10 (Oct. 2012), pp. 1270–1279. ISSN: 1538-3687. DOI: 10.1001/archneurol.2012.2052.

[35] Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature Genetics* 47.3 (Mar. 2015), pp. 291–295. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.3211.

7. The largest GWAS of longevity

A meta-analysis of genome-wide association studies identifies multiple longevity genes

Joris Deelen,* Daniel S. Evans,* Dan E. Arking, Niccolo' Tesi, Marianne Nygaard, Xiaomin Liu, Mary K. Wojczynski, Mary L. Biggs, Ashley van der Spek, Gil Atzmon, Erin B. Ware, Chloé Sarnowski *et al.*

* Authors contributed equally

This chapter was published in *Nature Communications*
https://www.nature.com/articles/s41467-019-11558-2

Abstract

Human longevity is heritable, but genome-wide association (GWA) studies have had limited success. Here, we perform two meta-analyses of GWA studies of a rigorous longevity phenotype definition including 11,262/3,484 cases surviving at or beyond the age corresponding to the $90^{th}/99^{th}$ survival percentile, respectively, and 25,483 controls whose age at death or at last contact was at or below the age corresponding to the 60^{th} survival percentile. Consistent with previous reports, *rs429358* (apolipoprotein E (ApoE) $\epsilon 4$) is associated with lower odds of surviving to the 90^{th} and 99^{th} percentile age, while *rs7412* (ApoE $\epsilon 2$) shows the opposite. Moreover, *rs7676745*, located near *GPR78*, associates with lower odds of surviving to the 90^{th} percentile age. Gene-level association analysis reveals a role for tissue-specific expression of multiple genes in longevity. Finally, genetic correlation of the longevity GWA results with that of several disease-related phenotypes points to a shared genetic architecture between health and longevity.

7.1 Background

The average human life expectancy has been increasing for centuries.[1] Based on twin studies, the heritability of human lifespan has been estimated to be ~25%, although this estimate differs among studies.[2] On the other hand, the heritability of lifespan based on the correlation of the mid-parent (*i.e.*, the average of the father and mother) and offspring difference between age at death and expected lifespan was estimated to be 12%.[3] A recent study has indicated that the different heritability estimates may be inflated due to assortative mating, leaving a true heritability that is below 10%.[4] The heritability of lifespan, estimated using the sibling relative risk, increases with age and is assumed to be enriched in long-lived families, particularly when belonging to the 10% longest-lived of their generation.[5, 2] To identify genetic associations with human lifespan, several genome-wide association (GWA) studies have been performed.[6, 2, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19] These studies have used a discrete (*i.e.*, older cases versus younger controls) or a continuous phenotype (such as age at death of individuals or their parents). The selection of cases for the studies using a discrete longevity phenotype has been based on the survival to ages above 90 or 100 years or belonging to the top 10% or 1% of survivors in a population. Studies defining cases using a discrete longevity phenotype often need to rely on controls from more contemporary birth cohorts, because all others from the case birth cohorts have died before sample collection. Previous GWA studies have identified several genetic variants, but the only locus that has shown genome-wide significance ($P \leq 5\text{x}10-8$) in multiple independent meta-analyses of GWA studies is apolipoprotein E (*APOE*),[20] where the ApoE $\epsilon 4$ variant is associated with lower odds of being a long-lived case. The lack of replication for many reported associations with longevity could be due, at least partly, to the use of different definitions for cases and controls between studies. Furthermore, even within a study, the use of a single age cut-off phenotype for men and women and for individuals belonging to different birth cohorts will give rise to heterogeneity, as survival probabilities differ by sex and birth cohort,[21] and genetic effects are known to be age- and birth cohort-specific.[5, 22] In an attempt to mitigate the effects of heterogeneous case and control groups, we use country-, sex- and birth cohort-specific life tables to identify ages that correspond to different survival percentiles to define cases and controls in our meta-analyses of GWA studies of longevity. Furthermore, most studies in our meta-analyses use controls from the same study population as the cases, which limits the impact of sampling biases that

could confound associations. The current meta-analyses include individuals from 20 cohorts from populations of European, East Asian, or African American descent. Two sets of cases are examined: individuals surviving at or beyond the age corresponding to the 90^{th} survival percentile (90^{th} percentile cases) or the 99^{th} survival percentile (99^{th} percentile cases) based on life tables specific to the country where each cohort was based, sex, and birth cohort (*i.e.*, birth year). The same country-, sex-, and birth cohort-specific life tables are used to define the age threshold for controls, corresponding to the 60^{th} percentile of survival. We identify two genome-wide significant loci, of which one is replicated in two independent European cohorts that use de novo genotyping. We also perform a gene-level association analysis based on tissue-specific gene expression and identify additional longevity genes. In addition, using linkage disequilibrium (LD) score regression,[23] we show that longevity is genetically correlated with multiple diseases and traits.

Table 7.1: **Samples included in the different genome-wide association meta-analyses or the replication and validation**

Study	Ancestry	90^{th} cases	99^{th} cases	All controls	Dead controls
Discovery					
100-plus/LASA/ADC	European	373	301	2271	245
AGES	European	300		1,001	466
CEPH[a]	European	1,234	1,112	831	
CHS	European	905	68	558	539
DKLS[a]	European	960	610	1,917	
FHS	European	332		1,444	539
GEHA Danish[a]	European	451	127	900	
GEHA French	European	271	81	358	
GEHA Italy	European	182		184	
HRS	European	361		3,312	657
LLFS	European	1,110	338	552	82
LLS + GEHA Dutch	European	1,037	377	712	
Longevity	European	548	271	584	
MrOS	European	1,171	82	386	320
Newcastle 85+[a]	European	215		5,159	
RS	European	774	79	2,965	1,731
SOF	European	812	37	354	300
Vitality 90+[a]	European	226		1,995	
Total		11,262	3,484	25,483	4,879
Replication					
DKLSII[a]	European	944	298	772	
GLS	European	1,613	1,613	4,215	
Total		2,557	1,911	4,987	
Validation					
UK Biobank	European	19,742	928	19,698	
Trans-ethnic					
CLHLS	East Asian	2,178	2,178	2,299	
CHS	African American	177		211	
Total		13,617	5,662	27,993	

100-plus: 100-plus Study; *LASA*, Longitudinal aging study of Amsterdam; *ADC*, Amsterdam dementia cohort; *AGES*, Age/Gene Environment Susceptibility Study; *CEPH*, CEPH centenarian cohort; *CHS*, Cardiovascular Health Study; *DKLS*, Danish longevity study; *FHS*, Framingham Heart Study; *GEHA*, Genetics of Healthy Aging Study; *HRS*, Health and Retirement Study; *LLFS*, Long Life Family Study; *LLS*, Leiden Longevity Study; *Longevity*, Longevity Gene Project; *MrOS*, Osteoporotic Fractures in Men Study; *Newcastle 85+*, Newcastle 85+ Study; *RS*, Rotterdam study; *SOF*, Study of Osteoporotic Fracture; *Vitality 90+*, Vitality 90+ project; *GLS*, German longevity study; *CLHLS*, Chinese Longitudinal Healthy Longevity Survey. [a] For these studies, controls were provided by a separate cohort. Further details of the cohorts are provided in Supplementary Data 4.

7.2 Results

7.2.1 Genome-wide association meta-analysis

We performed two meta-analyses in individuals of European ancestry combining cohort-specific genome-wide association data generated using 1000 Genomes imputation: (*i*) 90^{th} percentile cases versus all controls and (*ii*) 99^{th} percentile cases versus all controls. The numbers of cases and controls in each study are shown in Table 7.1. For both case definitions, multiple genetic variants at the well-replicated *APOE* locus reached genome-wide significance ($p \leq 5x10^{-8}$) (Table 7.2, Figure 7.1 and Figure 7.4). Consistent with previous reports, *rs429358* (ApoE $\epsilon 4$) was associated with lower odds of surviving to the 90^{th} or 99^{th} percentile age at the genome-wide significance level. In addition, we report a genome-wide significant association of *rs7412* (ApoE $\epsilon 2$) with higher odds of surviving to the 90^{th} and the 99^{th} percentile age. Conditional analysis in two of the cohorts with individuals of European ancestry, CEPH and LLS (combined with GEHA Dutch) (representing 18% of the 90^{th} percentile cases and 6% of all controls), indicated that the signal at the *APOE* locus was explained by these two independent variants, *i.e.*, *rs429358* (ApoE $\epsilon 4$) and *rs7412* (ApoE $\epsilon 2$). There was no evidence of heterogeneity of effect across cohorts for ApoE $\epsilon 2$ (p-value for heterogeneity (p_{het}) = 0.619, Table 7.2). For ApoE $\epsilon 4$, on the other hand, there was evidence of heterogeneity (p_{het} = 0.004, Table 7.2), although the direction of effect of this variant was consistent across cohorts (Figure 7.2). Besides ApoE $\epsilon 4$ and $\epsilon 2$, one additional variant, *rs7676745*, located on chromosome 4 near *GPR78*, showed a genome-wide significant association in the 90^{th} percentile cases versus all controls analysis (p= $4.3x10^{-8}$, Table 7.2). The rare allele of this variant (A) was associated with lower odds of surviving to the 90^{th} percentile age and there was no evidence of heterogeneity of effect across cohorts (Phet = 0.462, Table 7.2). The regional association and forest plots for this locus are depicted in Figure 7.1 and Figure 7.2. Most of the variants reported in Table 7.2 show stronger effects in the 99^{th} percentile as compared to the 90^{th} percentile analysis (Figure 7.5), indicating that the use of a more extreme phenotype results in stronger effects.

7.2.2 Replication

The effects of ApoE $\epsilon 4$ and $\epsilon 2$ were replicated in the two cohorts (*i.e.*, DKLSII and GLS) in which de novo genotyping, using predesigned Taqman SNP Genotyping Assays, was applied (Table 7.2). However, we were not able to replicate the effect of *rs7676745* in these cohorts, since there was no Taqman

Table 7.2: **Samples included in the different genome-wide association meta-analyses or the replication and validation**

rsID	Closest gene	Alleles	EAF	OR [95% CI]	P	I^2 (%)	P_{het}
90^{th} percentile cases versus all controls (Discovery)							
rs116362179	-	T/C	0.05	1.34 [1.20-1.50]	$4.9x10^{-7}$	0	0.457
rs7676745[a]	*GPR78*	A/G	0.04	0.67 [0.57-0.77]	$4.3x10^{-8}$	0	0.462
rs7754015	-	G/T	0.43	0.90 [0.86-0.94]	$6.8x10^{-7}$	0	0.670
rs35262860	*RP1*	GCT/G	0.39	1.11 [1.07-1.15]	$3.9x10^{-7}$	0	0.941
rs3138136	*RDH5*	T/C	0.10	0.83 [0.77-0.89]	$5.4x10^{-7}$	14.5	0.284
rs429358	*APOE*	C/T	0.13	0.60 [0.56-0.64]	$1.3x10^{-56}$	54.3	0.004
rs7412	*APOE*	T/C	0.09	1.28 [1.19-1.37]	$2.4x10^{-11}$	0	0.619
90^{th} percentile cases versus all controls (Replication)							
rs429358	*APOE*	C/T		0.45 [0.40-0.51]	$5.2x10^{-36}$	85.4	0.009
rs7412	*APOE*	T/C		1.32 [1.18-1.48]	$2.4x10^{-6}$	16.6	0.274
99^{th} percentile cases versus all controls (Discovery)							
rs3830412	*KALRN*	A/AT	0.22	1.21 [1.12-1.30]	$4.3x10^{-7}$	0	0.767
rs138762279	-	AT/A	0.16	0.79 [0.72-0.86]	$1.2x10^{-7}$	0	0.769
rs62502826	*KIF13B*	A/G	0.15	1.23 [1.23-1.33]	$5.6x10^{-7}$	14.9	0.298
rs7039467	*CDKN2A/B*	A/G	0.48	1.20 [1.12-1.28]	$1.1x10^{-7}$	0	0.843
rs429358	*APOE*	C/T	0.13	0.52 [0.47-0.58]	$3.9x10^{-34}$	0	0.833
rs7412	*APOE*	T/C	0.09	1.47 [1.32-1.64]	$3.2x10^{-12}$	0	0.639
99^{th} percentile cases versus all controls (Replication)							
rs429358	*APOE*	C/T		0.44 [0.38-0.50]	$4.0x10^{-32}$	84.0	0.012
rs7412	*APOE*	T/C		1.35 [1.19-1.53]	$2.0x10^{-6}$	0	0.534

Alleles, effect allele/other allele; *EAF*, effect allele frequency; *OR*, odds ratio (*i.e.*, odds to become long-lived when carrying the effect allele); *95% CI*, 95% confidence interval; I^2, heterogeneity statistic; p_{het}, *p*-value for heterogeneity; [a] We were not able to replicate the effect of this genetic variant, since there was no Taqman SNP Genotyping Assay available. We only report the most significant genetic variant for the loci with at least one variant with a *p*-value$\leq 1x10^{-6}$. The *rsID* is based on dbSNP build 150. The *Chr:Position* is based on Genome Reference Consortium Human Build 37 (GRCh37)

SNP Genotyping Assay available for this variant.

7.2.3 Validation in parental age-based data sets

Given that all available studies with genome-wide genetic data that met our inclusion criteria were included in our genome-wide association meta-analyses, we additionally set out to validate our findings in two UK Biobank parental longevity data sets (Table 7.1) and the parental lifespan data set recently created by Timmers and colleagues.[13] Since the genotyped individ-

uals in the UK Biobank were recruited at relatively young ages (40-69 years), these data sets were based on the age reached by the parents of the study participants. Hence, the phenotypes used for validation were different from those used in our meta-analyses, resulting in smaller effect sizes. Moreover, the reference panels used to impute the genetic variants (a merged panel of UK10K, 1000G Phase 3, and Haplotype Reference Consortium (HRC) for parental longevity and HRC alone for parental lifespan)[13] were different from the one used in our meta-analyses (1000G Phase 1), which could have influenced the outcome of the analyses. Of the variants that showed a p-value$\leq 1\text{x}10^{-6}$ in our meta-analyses (Table 7.2), only ApoE $\epsilon 4$ and $\epsilon 2$ were significantly associated with both parental longevity and lifespan ($p \leq 0.05$) in these data sets (Table 7.3). Moreover, the rare allele (A) of the second most significant variant at the *CDKN2A/B* locus, *rs2184061*, was associated with increased parental lifespan (p=8.4x10^{-6}), but not with parental longevity (p = 0.329). However, we had adequate power to validate all of our identified variants, even when the effect sizes were halved in the parental longevity data sets.

7.2.4 Trans-ethnic meta-analyses

We subsequently performed two trans-ethnic meta-analyses (90^{th} and 99^{th} percentile cases versus all controls) to see if the increase in sample size would lead to identification of additional longevity loci. In this analysis we included individuals of European (all previously used data sets), East Asian (CLHLS), and African American (CHS) ancestry. However, with the exception of *APOE* and *rs2069837*, located in *IL6*, which has previously been associated with longevity in CLHLS9, this analysis did not identify additional genome-wide significant loci (Table 7.4, Figure 7.3 and Figure 7.6). The observed association of the genetic variant in *IL6* in the trans-ethnic meta-analyses was mainly driven by the association in the East Asian population. The other variant previously associated with longevity in CLHLS9, *rs2440012*, located in *ANKRD20A9P*, did not pass quality control in the large majority of the included cohorts from populations of European descent and was thus not analysed in the trans-ethnic meta-analyses.

7.2.5 Comparison of control definitions

To examine the impact of the definition of controls, we performed a sensitivity analysis in which we compared the results of the meta-analysis using the same case definition (90^{th} percentile) with (*i*) all controls and (*ii*) dead controls

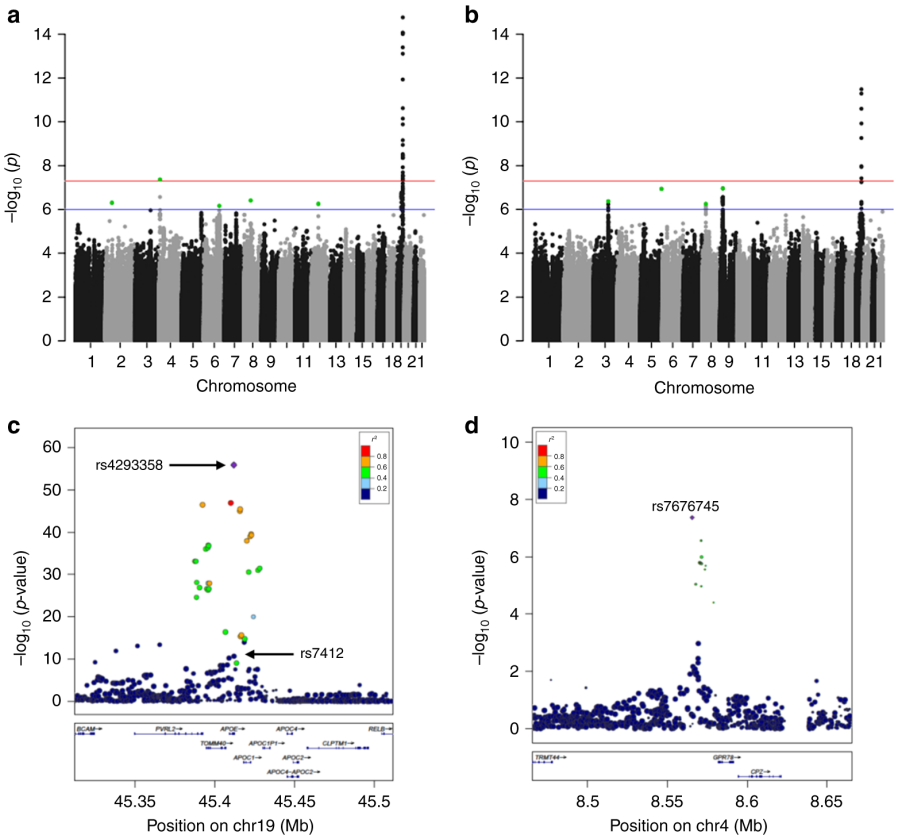


Figure 7.1: **Results of the European genome-wide association meta-analyses.** Manhattan plot presenting the $-\log_{10}$ *P-values* from the European genome-wide association meta-analysis of the 90^{th} percentile cases versus all controls (**a**) and 99^{th} percentile cases versus all controls (**b**). The red line indicates the threshold for genome-wide significance ($p \leq 5\text{x}10^{-8}$), while the blue line indicates the threshold for genetic variants that showed a suggestive significant association ($p \leq 1\text{x}10^{-6}$). The variants that are reported in Table 7.2 are highlighted in green. For representation purposes, the maximum of the y-axis was set to 14. Regional association plot for the *APOE* (**c**) and *GPR78* (**d**) loci based on the results from the 90^{th} percentile cases versus all controls meta-analysis. The colour of the variants is based on the linkage disequilibrium with *rs429358* (ApoE $\epsilon 4$) (**c**) or *rs7676745* (**d**)

only. For this analysis, only cohorts that contributed results using both control definitions were considered (*i.e.*, 100-plus/LASA/ADC, AGES, CHS, FHS, HRS, LLFS, MrOS, RS, and SOF). The results of the two meta-analyses with different control groups were very similar (Figure 7.7). Among the

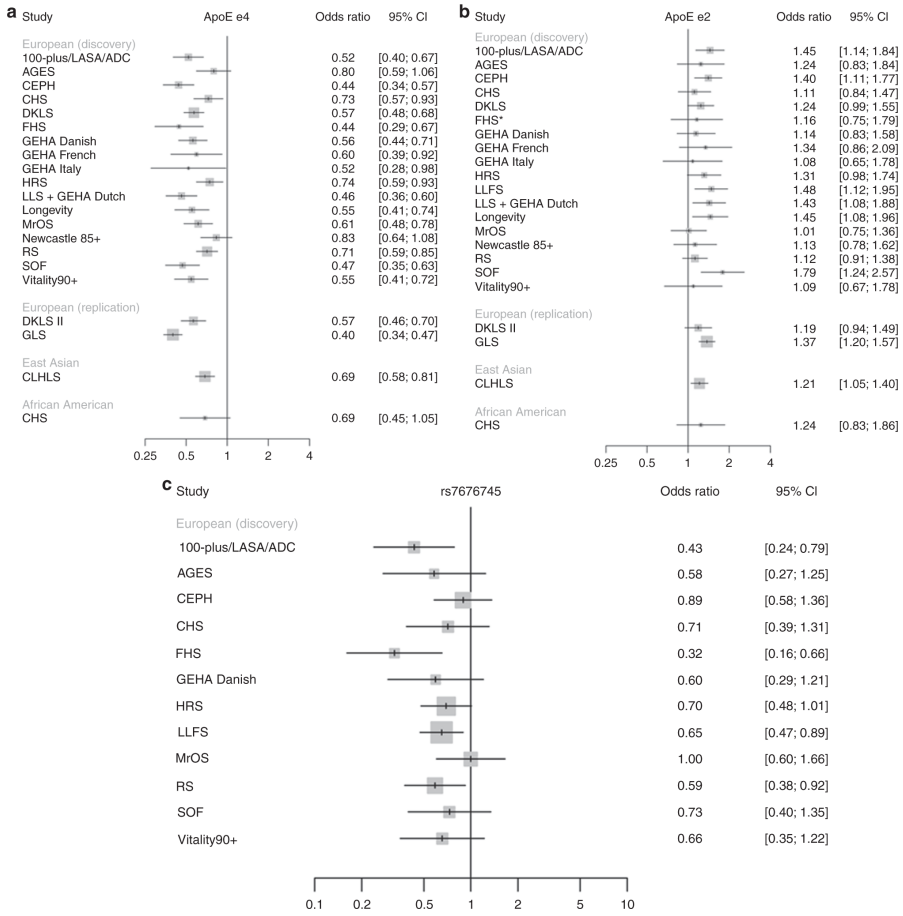


Figure 7.2: **Study-specific results for the genetic variants in *APOE* and *GPR78*.** Forest plots for the ApoE $\epsilon 4$ (**a**) and $\epsilon 2$ (**b**) variants and *rs7676745* (**c**) based on the results from the 90^{th} percentile versus all controls analysis. The size of the boxes represents the sample size of the cohort. We had no data available for ApoE $\epsilon 4$ in LLFS and for *rs7676745* in DKLS, GEHA Italy, GEHA Danish, LLS (combined with GEHA Dutch), Longevity, and Newcastle 85+. The data for ApoE $\epsilon 2$ in FHS was based on imputation using the Haplotype Reference Consortium reference panel due to the low-imputation quality of this variant when using the 1000 Genomes reference panel.

three loci with at least one genetic variant with a p-value$\leq 1\text{x}10^{-6}$ in either meta-analysis (and analysed in the same cohorts in both meta-analyses), the most significant variants had odds ratios (ORs) that differed by <1%

(Supplementary Table 1).

7.2.6 Replication of previously identified loci for human lifespan

To determine the association of previously identified loci for human lifespan and longevity, we performed a look-up of the reported genetic variants within these loci in our meta-analyses data sets. The only previously identified loci that contained variants that showed a significant (p<7.8x10^{-4}, *i.e.*, Bonferroni adjusted for the number of tested loci (N=64)) and directionally consistent associations in our study were *FOXO3* and *CDKN2A/B* (Supplementary Data 1). As depicted in Figure 7.8, the effects of the most frequently reported variants within these loci (*i.e.*, *rs2802292* and *rs1556516*) fluctuate between cohorts and there seems to be no correlation with the genetic background of the included populations. However, for the reported variants within both loci, the odds of surviving to the 99^{th} percentile age is higher than the odds of surviving to the 90^{th} percentile age, indicating they likely affect both early and late-life mortality. Several of the loci that have been associated with increased parental lifespan in the most recent and largest meta-analysis of GWA studies for this phenotype (*i.e.*, *KCNK3*, *HTT*, *LPA*, *ATXN2/BRAP*, and *LDLR*)[13] contain genetic variants that show a nominal significant association (p< 0.05) with higher odds of surviving to the 90^{th} and/or 99^{th} percentile age. Since the phenotypes used in our study (*i.e.*, cases surviving at or beyond the age corresponding to the 90^{th}/99^{th} survival percentile) were different from the one used in the previous study (*i.e.*, parental lifespan), we performed an additional look-up of these variants in one of the UK Biobank data sets we created for validation of our findings (*i.e.*, the 90^{th} percentile cases versus all controls data set). With the exception of the variant in *HTT*, all variants showed a nominal significant association in this data set (Supplementary Table 2), indicating that the lack of significant replication of these loci in our discovery phase data set is not likely to be due to a difference in the used phenotype.

7.2.7 Gene-level association analysis

In addition to genetic variant associations, GWA studies can also be used to identify gene-level associations by integrating results from *expression-quantitative-trait-locus* (eQTL) studies that relate variants to gene expression. In order to identify gene-level associations, we used *MetaXcan*, an analytic approach that uses tissue-specific eQTL results from the GTEx project to estimate gene-level associations with the trait examined from summary-level

Table 7.3: **Results of the validation in the UK Biobank parental age-based data sets**

rsID	Closest gene	Alleles	EAF	OR	95% CI	P
99^{th} percentile cases versus all controls (Parental longevity)						
rs116362179	-	T/C	0.04	1.01	0.94-1.08	0.775
rs7676745[a]	*GPR78*	A/G	0.04	0.98	0.92-1.06	0.667
rs7754015	-	G/T	0.43	1.00	0.97-1.03	0.832
rs35262860	*RP1*	GCT/G	0.39	0.97	0.94-0.99	0.021
rs3138136	*RDH5*	T/C	0.11	1.00	0.95-1.04	0.863
rs429358	*APOE*	C/T	0.16	0.85	0.81-0.88	$1.1x10^{-16}$
rs7412	*APOE*	T/C	0.08	1.12	1.06-1.18	$2.2x10^{-5}$
90^{th} percentile cases versus all controls (Parental longevity)						
rs116362179	-	T/C	0.04	1.00	0.98-1.02	0.697
rs7676745[a]	*GPR78*	A/G	0.05	1.01	0.99-1.03	0.247
rs3138136	*RDH5*	T/C	0.11	0.99	0.98-1.00	0.135
rs429358	*APOE*	C/T	0.15	0.90	0.89-0.91	$3.1x10^{-83}$
rs7412	*APOE*	T/C	0.08	1.06	1.05-1.08	$7.6x10^{-17}$
99^{th} percentile cases versus all controls (Parental longevity)						
rs3830412	*KALRN*	A/AT	0.20	1.11	0.99-1.24	0.081
rs138762279	-	AT/A	0.34	1.05	0.95-1.17	0.299
rs62502826	*KIF13B*	A/G	0.14	1.04	0.90-1.19	0.614
rs7039467	*CDKN2A/B*	A/G	0.69	0.93	0.83-1.05	0.245
rs2184061	*CDKN2A/B*	A/C	0.40	0.95	0.87-1.05	0.329
rs429358	*APOE*	C/T	0.16	0.76	0.66-0.87	$9.6x10^{-5}$
rs7412	*APOE*	T/C	0.08	1.23	1.05-1.45	0.011
99^{th} percentile cases versus all controls (Parental longevity)						
rs62502826	*KIF13B*	A/G	0.14	1.00	0.99-1.02	0.376
rs2184061	*CDKN2A/B*	A/C	0.40	1.02	1.01-1.03	$8.4x10^{-6}$
rs429358	*APOE*	C/T	0.15	0.90	0.89-0.91	$3.1x10^{-84}$
rs7412	*APOE*	T/C	0.08	1.06	1.05-1.08	$7.6x10^{-17}$

For the *CDKN2A/B* locus we have also reported the second most significant variant in this locus (*rs2184061*), since the allele frequency of the most significant variant (*rs7039467*) is not comparable between the meta-analyses and UK Biobank data sets due to difference in the reference panel used for imputation. The *rsID* is based on dbSNP build 150. The *Chr:Position* is based on Genome Reference Consortium Human Build 37 (GRCh37). *Alleles*, effect allele/other allele; *EAF*, effect allele frequency; *OR [95% CI]*, odds ratio (*i.e.*, odds of parent(s) to become long-lived when carrying the effect allele), and relative 95% confidence interval.

GWA study results.[24] Tissue-specific genetically predicted expression of 14 genes (*ANKRD31, BLOC1S1, KANSL1, CRHR1, ARL17A, LRRC37A2, ERCC1, RELB, DMPK, CD3EAP, PVRL2, GEMIN7, BLOC1S3*, and *APOC2*) was significantly associated with survival to the 90^{th} and/or 99^{th} percentile age after adjustment for multiple testing (Table 7.5). Eight of these genes (*ERCC1, RELB, DMPK, CD3EAP, PVRL2, GEMIN7, BLOC1S3*, and *APOC2*) are located near the *APOE* gene, raising the likely possibility that these associations reflected the influence of variants in this well-established longevity-associated locus. The remaining genes are located on chromosome 5, 12, and 17. As depicted in Supplementary Data 2, distinct sets of genetic variants were used by *MetaXcan* for all significant tissue-specific gene expression associations with survival to the 90^{th} and/or 99^{th} percentile age.

7.2.8 Genetic correlation analyses

LD score regression was performed to determine the genetic correlation between the different case definitions used for our meta-analyses (based on the results from the European cohorts only), and between longevity and other traits and diseases.[23] The genetic correlation (rg) between the 90^{th} and 99^{th} percentile analysis, using all controls for both groups, was 1.01 (*SE* = 0.06, $p = 3.9\text{x}10^{-66}$). Using LD Hub,[25] which performs automated LD score regression, we subsequently estimated the genetic correlation of our phenotypes with 246 diseases and traits available in their database. We found a significant genetic correlation of our phenotypes with the father's age at death phenotype from the UK Biobank. The most significant (negative) genetic correlation of both our phenotypes was with coronary artery disease (CAD) (*rg* (SE) = -0.40 (0.07) and *rg* (SE) = -0.29 (0.07), respectively) and several traits involved in type 2 diabetes (T2D) also showed a significant association with one or both phenotypes after Bonferroni adjustment for multiple testing (Table 7.6 and Supplementary Data 3).

Table 7.4: **Results of the trans-ethnic genome-wide association meta-analyses**

rsID	Closest gene	Alleles	EAF	OR [95% CI]	P	I^2 (%)	p_{het}
90^{th} percentile cases versus all controls							
rs12143832	*ECE1*	C/T	0.46	0.90 [0.87-0.94]	$2.0x10^{-7}$	0	0.722
rs7676745[a]	*GPR78*	A/G	0.04	0.67 [0.58-0.78]	$1.7x10^{-7}$	1.8	0.428
rs1262476	-	A/G	0.24	1.12 [1.07-1.17]	$9.8x10^{-7}$	0	0.574
rs2069837	*IL6*	G/A	0.08	0.90 [0.82-0.99]	$5.2x10^{-8}$	50.7	0.005
rs35262860	*RP1*	GCT/G	0.39	1.11 [1.07-1.15]	$5.6x10^{-7}$	0	0.955
rs62127362	*CEP89*	C/G	0.13	0.87 [0.82-0.93]	4.3x10−7	21.4	0.190
rs429358	*APOE*	C/T	0.13	0.60 [0.55-0.66]	$1.0x10^{-61}$	52.1	0.004
rs7412	*APOE*	T/C	0.09	1.26 [1.19-1.35]	$1.7x10^{-12}$	0	0.718
99^{th} percentile cases versus all controls							
rs2758603	*PMF1*	C/T	0.34	1.12 [1.02-1.22]	$9.8x10^{-7}$	57.2	0.005
rs3830412	*KALRN*	A/AT	0.22	1.21 [1.12-1.30]	$8.2x10^{-7}$	0	0.767
rs138762279	-	AT/A	0.16	0.79 [0.72-0.86]	$2.2x10^{-7}$	0	0.769
rs2069837	*IL6*	G/A	0.09	0.90 [0.76-1.08]	$1.4x10^{-8}$	67.7	$3.5x10^{-4}$
rs7039467	*CDKN2A/B*	A/G	0.48	1.20 [1.12-1.28]	$2.1x10^{-7}$	0	0.843
rs429358	*APOE*	C/T	0.13	0.55 [0.50-0.61]	$1.3x10^{-36}$	20.0	0.247
rs7412	*APOE*	T/C	0.09	1.39 [1.26-1.53]	$1.7x10^{-12}$	10.0	0.347

We only report the most significant genetic variant for the loci with at least one variant with a p-value $\leq 1x10^{-6}$. The reported p is the p-value from the Han-Eskin random-effects (RE^2) model from *METASOFT*. The *rsID* is based on dbSNP build 150; the *Chr:Position* is based on Genome Reference Consortium Human Build 37 (GRCh37); *Alleles*, effect allele/other allele; *EAF*, effect allele frequency (based on individuals of European ancestry only); *OR*, odds ratio (*i.e.*, odds to become long-lived when carrying the effect allele); *95% CI*, 95% confidence interval; I^2, heterogeneity statistic; p_{het}, p-value for heterogeneity.

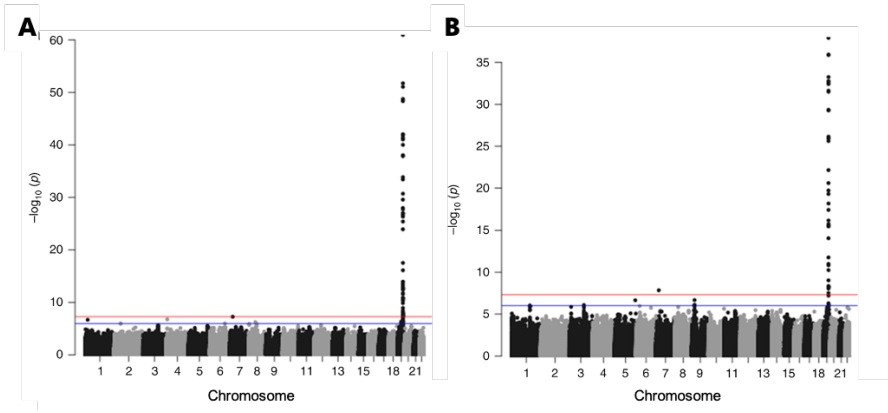


Figure 7.3: **Results of the trans-ethnic genome-wide association meta-analyses.** Manhattan plot presenting the $-\log_{10}$ p-values from the trans-ethnic genome-wide association meta-analysis of the 90^{th} percentile cases versus all controls (**A**) and 99^{th} percentile cases versus all controls (**B**). The red line indicates the threshold for genome-wide significance ($p \leq 5\text{x}10^{-8}$), while the blue line indicates the threshold for genetic variants that showed a suggestive significant association ($p \leq 1\text{x}10^{-6}$).

7.3 Discussion

We brought together studies from all over the world to perform GWA study meta-analyses in over 13,000 long-lived individuals of diverse ethnic background, including European, East Asian and African American ancestry, to characterise the genetic architecture of human longevity. We used the 1000 Genomes reference panel for imputation to expand the coverage of the genome in comparison to previous GWA studies of longevity. Consistent with previous reports, *rs429358*, defining ApoE $\epsilon 4$, was associated with decreased odds of becoming long-lived. Moreover, we report a genome-wide significant association of *rs7412*, defining ApoE $\epsilon 2$, with increased odds of becoming long-lived. We additionally found a genome-wide significant association of a locus near *GPR78*. Gene-level association analysis revealed association of increased *KANSL1, CRHR1, ARL17A,* and *LRRC37A2* expression and decreased *ANKRD31* and *BLOC1S1* expression with increased odds of becoming long-lived. Genetic correlation analysis showed that our longevity phenotypes are genetically correlated with father's age at death, CAD and T2D-related phenotypes. Genetic variation in *APOE* is well known to be associated with longevity and lifespan, with the first report more than two dec- ades ago

in a small candidate gene study.[26] Since then, there have been numerous candidate gene studies, including individuals of diverse ancestry, which have identified associations of ApoE with longevity.[27, 28, 29, 30] However, thus far, *rs7412*, the ApoE $\epsilon 2$-defining, genetic variant has not been reported to show a genome-wide significant association in GWA studies of longevity and lifespan. This could be due to the fact that we performed imputation using the 1000 Genomes reference panel, while earlier GWA studies used the HapMap reference panel, which has limited coverage of this variant. ApoE mediates cholesterol metabolism in peripheral tissues and is the principal cholesterol carrier in the brain. The ApoE $\epsilon 2$ and $\epsilon 4$ variants have previously been associated with a decreased ($\epsilon 2$) or increased ($\epsilon 4$) risk for several age-related diseases, such as cardiovascular disease and Alzheimer's disease,[31] which could explain their effect on longevity. The fact that the two variants in ApoE show opposite effects may be attributable to differences in structural and biophysical properties of the protein, since ApoE $\epsilon 2$ shows high stability and ApoE $\epsilon 4$ low stability upon folding.[32] We also found a genome-wide significant association of *rs7676745*, located on chromosome 4 near *GPR78*. We have to note that this locus would benefit from replication in independent cohorts in the future, given that we were not able to replicate this variant in the cohorts in which de novo genotyping was applied. There is no report of association of this locus with other traits according to Phenoscanner (`http://www.phenoscanner.medschl.cam.ac.uk/`),[33] although other genetic variants in this gene have been associated with several diseases and traits in the UK Biobank, including death due to a variety of disorders. The *GPR78* protein, belongs to the family of G-protein-coupled receptors, whose main function is to mediate physiological responses to various extracellular signals, including hormones and neuro-transmitters.[34] However, the specific function of *GPR78* is still largely unknown, although it has been shown to play a role in lung cancer metastasis.[35] To maximise power for discovery, we meta-analysed results from all of the studies that contained long-lived individuals that met our 90^{th} and/or 99^{th} percentile case definitions, had genome-wide genetic data, and were able to participate. Hence, we were not able to replicate our findings in an independent cohort with genome-wide genotype data and participants reaching the age of our case definitions. Therefore, we tried to validate our findings using two related phenotypes, parental longevity and lifespan, in the UK Biobank. We applied our case and control definitions to the parental lifespan of genotyped middle-aged UK Biobank participants rather than the participants themselves, as none of the latter fulfilled the age criteria for cases in our study. Although this

resulted in relatively large data sets for both the 90^{th} and 99^{th} percentile analysis, the power to replicate our findings using the parental longevity traits was lower in comparison to replication using the traits based on the genotyped individuals themselves, since these individuals share only half of their parental genomes. In addition, many of the genotyped individuals, who were 40-69 years at recruitment, will never reach the age belonging to the 90^{th}, let alone the 99^{th}, percentile of their birth cohort. This may explain why we were unable to validate any of our suggestive associations ($p \leq 1x10^{-6}$), with the exception of the genetic variants at the *APOE* locus in these data sets. On the other hand, we were able to validate one additional locus, *CDKN2A/B*, in the parental lifespan data set. This is not surprising, since this locus had already been reported to associate with parental lifespan.[13] However, it is unclear why our reported variants at this locus, *rs7039467* and *rs2184061*, are not associated with parental longevity, given that the most significant parental lifespan-associated variant at this locus, *rs1556516*, also shows a nominal significant effect on parental longevity (see Supplementary Table 2). We hypothesise that this may be due to a difference in the LD structure of the reference panels used for imputation.

We were able to detect significant genetic associations at two previously identified longevity-related loci, *FOXO3* and *CDKN2A/B*. For the other loci, we did not find evidence for replication ($p > 7.8x10^{-4}$), despite having adequate power (≥0.8) for replication of all but one of the examined genetic variants (*rs28926173*) associated with the discrete longevity phenotypes. We were not able to calculate our power to replicate the variants associated with the continuous lifespan-related phenotypes, although we should have had adequate power to replicate variants with a minor allele frequency (MAF) > 12% and an OR > 1.1 (based on the 90^{th} percentile versus all controls analysis). However, several of the variants associated with parental lifespan show a directionally consistent and nominal significant association with our phenotypes, indicating they may also be relevant for longevity. The failure to replicate previously reported loci could be due to the use of a different longevity phenotype then what was used in previous studies, the small effect size of some of the variants associated with parental lifespan, and the modest power of our study. The fact that we detect significant associations of variants in the *FOXO3* locus is not surprising, since this locus was previously reported in the longevity GWA study from the CHARGE consortium,[6] from which many cohorts are included in these meta-analyses. So far, three functional longevity-associated variants have been identified at the *FOXO3* locus (*rs2802292, rs12206094*, and *rs4946935*). For all of them,

Table 7.5: **Results of the trans-ethnic genome-wide association meta-analyses**

Genes	Tissue	OR_{90}	p_{90}	OR_{99}	p_{99}
ANKRD31	Stomach	0.63	$\mathbf{1.1x10^{-6}}$	0.61	$9.0x10^{-4}$
BLOC1S1	Adipose subcutaneous	0.49	$\mathbf{4.5x10^{-7}}$	0.56	0.009
KANSL1	Skin sun exposed lower leg	1.22	$\mathbf{1.5x10^{-6}}$	1.26	$1.9x10^{-4}$
CRHR1	Nerve tibial	1.54	$\mathbf{3.4x10^{-7}}$	1.81	$6.2x10^{-6}$
ARL17A	Artery aorta	1.24	$\mathbf{8.1x10^{-7}}$	1.31	$5.9x10^{-5}$
ARL17A	Breast mammary tissue	1.18	$\mathbf{1.8x10^{-6}}$	1.22	$3.2x10^{-4}$
ARL17A	Colon sigmoid	1.21	$\mathbf{2.2x10^{-6}}$	1.21	0.002
LRRC37A2	Minor salivary gland	1.17	$\mathbf{2.2x10^{-6}}$	1.20	$4.4x10^{-4}$
ERCC1	Ovary	1.19	$\mathbf{2.8x10^{-6}}$	1.24	$1.8x10^{-4}$
RELB	Lung	0.57	$\mathbf{2.0x10^{-7}}$	0.44	$2.9x10^{-6}$
DMPK	Stomach	1.64	$\mathbf{1.7x10^{-6}}$	2.31	$1.8x10^{-6}$
CD3EAP	Brain substantia nigra	0.51	$\mathbf{8.0x10^{-17}}$	0.36	$\mathbf{3.8x10^{-15}}$
PVRL2	Artery coronary	1.36	$\mathbf{5.0x10^{-7}}$	1.59	$1.6x10^{-6}$
PVRL2	Oesophagus muscularis	1.62	$\mathbf{6.6x10^{-7}}$	2.31	$\mathbf{4.4x10^{-8}}$
GEMIN7	Brain nucleus accumbens basal ganglia	0.85	$1.5x10^{-4}$	0.70	$\mathbf{1.4x10^{-7}}$
BLOC1S3	Oesophagus muscularis	2.80	$\mathbf{6.4x10^{-16}}$	4.47	$\mathbf{1.3x10^{-13}}$
APOC2	Skin not sun exposed suprapubic	0.75	$\mathbf{4.2x10^{-7}}$	0.74	$9.3x10^{-4}$

OR, odds ratio (*i.e.*, odds to become long-lived when having an increased tissue-specific gene expression); *p*-values, highlighted in bold are significant after adjustment for multiple testing of 247,999 longevity associations with gene-tissue pairs (Storey *q*-value<0.05); OR_{90} and p_{90} are based on the analysis of the 90^{th} percentile cases versus all controls meta-analysis data set, while OR_{99} and p_{99} are based on the analysis of the 99^{th} percentile cases versus all controls meta-analysis data set

an allele-specific response to cellular stress was observed. Consistently, the longevity-associated alleles of all three variants were shown to induce *FOXO3* expression.[36, 37] The *CDKN2A/B* locus has previously been associated with parental lifespan and parents' attained age in the UK Biobank as well as a diversity of age-related diseases.[38, 13, 39] The longevity-associated allele of the most significant variant at this locus (*rs1556516*) has also been associated with lower odds of developing CAD.[40] Although the molecular mechanism behind this association is still unclear, it is known that genes encoded at the *CDKN2A/B* locus are involved in cellular senescence,[41] a known hallmark of ageing in animal models.[42] The gene-level association analysis identified several associations between increased (*KANSL1, CRHR1, ARL17A*, and *LRRC37A2*) or decreased (*ANKRD31* and *BLOC1S1*) genetically driven tissue-specific gene expression with survival to the 90^{th} percentile age. The increased expression of *KANSL1, CRHR1, ARL17A*, and *LRRC37A2* on chromosome 17q21.31 is regulated by different genetic variants, indicating that these associations may be independent. More functional work is needed to determine the exact relationship between the altered genetically driven tissue-specific expression of these genes and longevity in humans. A limitation of *MetaXcan* is that the underlying GTEx models might not have been adequately adjusted for age, which could be problematic for an age-related phenotype like longevity. However, *MetaXcan* has successfully been used to identify gene-level associations with age-related diseases and traits, such as Alzheimer's disease and age-related macular degeneration.[24] The genetic correlation analyses showed that survival to ages corresponding to the 90^{th} and 99^{th} percentile shared genetic associations with father's age at death, CAD and T2D-related phenotypes, suggesting that survival to old ages may at least partially be explained by protective influences on the mechanisms underlying these traits. The genetic correlation with CAD and T2D-related phenotypes is expected, since it has previously been reported that individuals from long-lived families show a decreased prevalence of cardiovascular disease and T2D.[43, 44] The higher genetic correlation of our longevity phenotypes with father's in comparison to mother's age at death may be explained by the difference in the prevalence of cardiovascular diseases and T2D between men and women in the last century,[45, 46] which may be, at least partially, attributable to a difference in smoking prevalence.[47] Hence, the correlation of our longevity phenotypes with the parental age at death phenotypes from UK Biobank is likely due to the absence of death from specific diseases (*i.e.*, those with a higher prevalence in men). For longevity-specific loci, on the other hand, one would expect

that they will have beneficial effects on multiple diseases simultaneously, since long-lived individuals show a delay in overall morbidity.[48] Our study design imposed an age gap between cases and controls to reduce outcome misclassification, which we expected could potentially increase power by increasing the genetic effect size. It has been correctly noted that longevity study designs that include an age gap between cases and controls result in an effect estimate that is based on an OR and a relative risk (RR) term, which could lead to the identification of genetic variant associations related to early mortality (OR), rather than survival past the case age threshold (RR) (for more details see Sebastiani *et al.*).[49] However, we have presented evidence that imposing a case-control age gap did not greatly influence our results or prevent our replication of variant associations previously discovered using study designs without a case-control age gap. First, our sensitivity analysis indicated that reducing the age gap between cases and controls had a minimal effect on our results. Our sensitivity analysis compared results using dead controls, where all individuals had died before they reached the 60^{th} percentile age, and all controls, which included dead controls and individuals whose age at last contact was below the 60^{th} percentile age but whose age of death was unknown. There is likely to be some outcome misclassification of the living controls, since a small percentage may survive beyond the age corresponding to the 90^{th} or 99^{th} survival percentile. On the other hand, the age gap between cases and controls was narrower for all controls compared to dead controls. However, despite the narrower age gap, the suggestively significant results in all controls and dead controls comparisons with 90^{th} percentile cases were essentially unchanged, and there was a very high genetic correlation between the results of these two meta-analyses, indicating that the age gap had little or no impact on our results. Second, if we had discovered a large number of genome-wide significant variant associations in our study, it could be argued that the OR, reflecting early mortality, contributed to some or all of them. However, the only genome-wide significant variant associations we detected were in the *APOE* locus, which have been identified using multiple study designs, including designs with no prespecified age gap between cases and controls,[12] and the *GPR78* locus. Third, it is unlikely that our study design prevented the replication of findings from previous GWA studies of survival to extreme ages (*i.e.*, 99^{th} percentile cases) that did not include a case-control age gap, since such studies would only identify variants associated with survival past the minimum case age and not with early mortality. For variants with no early mortality association, it would be expected that the association estimate in our study would have an OR equal

Table 7.6: **Results of the genetic correlation analyses of the 90^{th} and 99^{th} percentile phenotypes with other diseases and traits**

Disease/Trait	rg_{90} (SE_{90})	p_{90}	rg_{99} (SE_{99})	p_{99}
Coronary artery disease	-0.40 (0.07)	$\mathbf{1.7x10^{-8}}$	-0.29 (0.07)	$\mathbf{1.2x10^{-5}}$
Fathers age at death	0.74 (0.13)	$\mathbf{2.5x10^{-8}}$	0.54 (0.13)	$\mathbf{2.7x10^{-5}}$
HDL cholesterol	0.36 (0.07)	$\mathbf{1.0x10^{-7}}$	0.22 (0.07)	0.002
Age of first birth	0.33 (0.07)	$\mathbf{3.8x10^{-7}}$	0.16 (0.07)	0.019
Years of schooling 2016	0.26 (0.05)	$\mathbf{9.6x10^{-7}}$	0.12 (0.05)	0.017
Waist circumference	-0.26 (0.05)	$\mathbf{2.4x10^{-6}}$	-0.19 (0.06)	0.001
Type 2 diabetes	-0.44 (0.10)	$\mathbf{4.4x10^{-6}}$	-0.42 (0.10)	$\mathbf{2.0x10^{-5}}$
Overweight	-0.28 (0.06)	$\mathbf{1.2x10^{-5}}$	-0.23 (0.07)	$9.0x10^{-4}$
Fastin insulin main effect	-0.45 (0.11)	$\mathbf{3.0x10^{-5}}$	-0.33 (0.11)	0.002
Urate	-0.26 (0.07)	$\mathbf{5.0x10^{-5}}$	-0.15 (0.06)	0.013
Body mass index	-0.21 (0.05)	$\mathbf{9.2x10^{-5}}$	-0.19 (0.07)	0.004
Cigarettes smoked per day	-0.49 (0.13)	$\mathbf{1.0x10^{-4}}$	-0.31 (0.13)	0.016
Mothers age at death	0.51 (0.14)	$\mathbf{2.0x10^{-4}}$	0.14 (0.13)	0.289
Waist-to-hip ratio	-0.24 (0.07)	$\mathbf{2.0x10^{-4}}$	-0.15 (0.07)	0.028

p-values highlighted in bold are significant after Bonferroni adjustment for multiple testing ($p < 0.05/246$). rg_{90}, SE_{90}, and p_{90} are based on the analysis of the 90^{th} percentile cases versus all controls meta-analysis data set, while rg_{99}, SE_{99}, and p_{99} are based on the analysis of the 99^{th} percentile cases versus all controls meta-analysis data set. *rg*, genetic correlation; *SE*, standard error of the *rg* estimate; *HDL*, high-density lipoprotein.

to one and a RR greater than one. Nothing prevents our study design from also detecting this type of variant association, as our estimated association parameter reflects both the OR and RR.

The majority of the previously performed GWA studies of longevity used the survival of individuals to a pre-defined age threshold (*i.e.*, 85, 90, or 100 years) as selection criterion to define long-lived cases. Although these studies used a consistent phenotype for each cohort included in the GWA study, this type of selection may gave rise to heterogeneity, given that survival probabilities differ between sexes and birth cohorts.[21] Moreover, it was recently shown that the heritable component of longevity is strongest in individuals belonging to the top 10% survivors of their birth cohort.[2] Hence, instead of using a pre-defined age threshold, we decided to select cases based on country-, sex- and birth cohort-specific life tables. For the definition of controls we used the 60^{th} percentile age, since we wanted to include as many controls as possible (preferably from the same cohort as

the cases), while leaving a large enough age gap between our cases and controls. Using the 1920 birth cohort as an example, the difference between the 60^{th} and 90^{th} percentile age is 14 years (men) or 11 years (women), which is quite substantial. The difference between the 70^{th} and 90^{th} percentile age, on the other hand, is considerably smaller (9 years (men) or 7 years (women)) and the living controls are more likely to reach the 90^{th} percentile age, which increases the risk of outcome misclassification. Moreover, even when selecting the 60^{th} percentile controls from much later birth cohorts (*i.e.*, 1940) than the cases (*i.e.*, 1900) the ages will not overlap. Our study has several limitations. First, we did not analyse the sex and mitochondrial chromosomes, since we were unable to gather enough cohorts that could contribute to the analysis of these chromosomes. However, these chromosomes may harbour loci associated with longevity that we thus have missed. Second, although we included as many cohorts as possible, the sample size of our study is still relatively small (especially for the 99^{th} percentile analysis) in comparison to GWA studies of age-related diseases, such as T2D and cardiovascular disease, and parental age at death.[9, 50, 51] Hence, this limited our power to detect loci with a low MAF (<1%) that contribute to longevity. Third, we did not perform sex-stratified analyses and may thus have missed sex-specific longevity-related genetic variants. The reason for this is that (*i*) we only identified a limited number of suggestive significant associations in our unstratified 90^{th} and 99^{th} percentile analyses, (*ii*) our sample size is modest (especially when stratified by sex), and (*iii*) thus far, there has been no report of any genome-wide significant sex-specific longevity locus. Given that we have included nearly all cohorts with long-lived individuals with genome-wide genetic data in our study, it will be challenging to increase the sample size in future GWA studies using the same extreme phenotypes. Future genetic studies of longevity may therefore benefit from the use of alternative phe- notypes or more rigorous phenotype definitions. Alternative phenotypes that could be used are the parental lifespan or healthspan-related phenotypes that were analysed in the UK Biobank or biomarkers of healthy aging.[13, 52, 53] One way to strengthen the longevity phenotype is by selecting cases from families with multiple individuals belonging to the top 10% survivors of their birth cohort.[2] Moreover, given the limited number of longevity-associated genetic variants identified through GWA studies and the availability of affordable exome and whole-genome sequencing, future genetic studies of longevity may also benefit from the analysis of rare genetic variants. Ideally, such studies should also try to include participants from genetically diverse populations. Most cohorts that are currently included in

genetic longevity studies originate from populations of European descent, while some longevity loci may be specific for non-European populations, as exemplified by the previously reported genome-wide associations of genetic variants in *IL6* and *ANKRD20A9P* in Han Chinese.[8] Moreover, a recent genetic study of multiple complex traits has shown the benefit of analysis of diverse populations.[54] In conclusion, we performed a genome-wide association study of longevity-related phenotypes in individuals of European, East Asian and African American ancestry and identified the *APOE* and *GPR78* loci to be associated with these phenotypes in our study. Moreover, our gene-level association analyses highlight a role for tissue-specific expression of genes at chromosome 5q13.3, 12q13.2, 17q21.31, and 19q13.32 in longevity. Genetic correlation analyses show that our longevity-related phenotypes are genetically correlated with several disease-related phenotypes, which in turn could help to identify phenotypes that could be used as potential biomarkers for longevity in future (genetic) studies.

7.4 Methods

7.4.1 Study populations

In this collaborative effort, we included cohorts that participated in one or more of the previously published GWA studies on longevity.[6, 7, 8] The sample sizes and descriptive characteristics of the cohorts used in this study are provided in Table 7.1, Supplementary Data 4, and the Supplementary Methods. We have complied with all relevant ethical regulations for work with human subjects. All participants provided written informed consent and the studies were approved by the relevant institutional review boards.

7.4.2 Case and control definitions

Cases were individuals who lived to an age above the 90^{th} or 99^{th} percentile based on cohort life tables from census data from the appropriate country, sex, and birth cohort. Controls were individuals who died at or before the age at the 60^{th} percentile or whose age at the last follow-up visit was at or before the 60th percentile age. Hence, the number of selected cases and controls is defined by the ages of their birth cohort corresponding to the 60^{th} or $90^{th}/99^{th}$ percentile age and is independent of the study population used (*i.e.*, the number of controls and cases within a study population is not based on the percentiles of that specific population, but instead on that of their birth cohorts). As part of their recruitment protocol, many of the studies enroled participants that were already relatively old at the time of recruitment (*i.e.*, close to (or even over) the 60^{th} percentile age). The majority of these individuals subsequently survived past the 60^{th} percentile age threshold of their respective birth cohorts, resulting in a small number of controls in comparison to the number of cases for some of these studies. The cohort life tables were available through the Human Mortality Database (`www.mortality.org`),[55] the United States Social Security Administration (here)[21] or National registries (here). For example, the 60^{th}, 90^{th}, and 99^{th} percentile correspond to ages of 75, 89, and 98 years for men and 83, 94, and 102 years for women for the 1920 birth cohort from the US. For cohort life tables providing birth cohort by decade, linear model predictions were used to estimate the ages corresponding to survival percentiles at yearly birth cohorts. For the parental longevity analyses in the UK Biobank, cases were individuals with at least one parent achieving an age above the 90^{th} or 99^{th} percentile and who had not themselves died, while controls were individuals for whom both parents died at or before the age at the 60^{th} percentile.

7.4.3 Genome-wide association analysis of individual cohorts

Details on the genotyping (platform and quality control criteria), imputation and genome-wide association analyses for each cohort are provided in Supplementary Data 5. In all cohorts, genetic variants were imputed using the 1000G Phase 1 version 3 reference panel. The logistic regression analyses were adjusted for clinical site, known family relationships, and/or the first four principal components (if applicable). All cohorts used a Hardy-Weinberg equilibrium (HWE) p-value that was between $1x10^{-4}$ and $1x10^{-6}$ to exclude variants not in HWE, which is considered standard in GWA studies. However, this may have resulted in removal of variants that were out of HWE in the cases due to mortality selection.[56]

7.4.4 Quality control of individual cohorts

Quality control of the summary statistics from each cohort was performed using the *EasyQC* software and the standard script (fileqc-1000G.ecf) available on their website.[57] The only difference was that we used the expected minor allele count (eMAC) instead of the MAC. To this end, we first calculated the *Effective N* ($2/(1/N_{cases} + 1/N_{controls})$) for each cohort. The use of the *Effective N* instead of the *Total N* leads to a more stringent filtering of genetic variants and decreases the chance of false positive findings due to an imbalance between the number of cases and controls.[57] The *Effective N* was subsequently used to calculate the eMAC (2 x minor allele frequency x *Effective N* x imputation quality) for each variant. Variants were excluded when eMAC < 10, with the exception of the Newcastle 85 + (90^{th} percentile cases versus all controls) and the RS (99^{th} percentile cases versus all controls) data sets in which we excluded variants when eMAC < 25 due to the large imbalance between the number of cases and controls in these data sets (1:24 and 1:38, respectively) in comparison to the other ones (all < 1:10). For the CLHLS and LLFS data sets, we flipped the strands of several variants based on the dis- cordance of allele frequencies with the reference panel. We only flipped palindromic variants with a MAF < 0.4 and an allele frequency that differed from the reference panel by <10% after switching.

7.4.5 Meta-analyses

The fixed-effect meta-analyses based on the data sets with individuals of European ancestry were performed on the cleaned files using *METAL*,[58] with the *Effective N* as weight and adjustment for genomic control (lambda (λ)) for each cohort. Cohorts with an *Effective N* < 50 were excluded from

the meta- analyses. We did not apply genomic control on the meta-analyses results, since there was limited inflation (all λ < 1.04, Figure 7.4). The trans-ethnic meta-analyses were performed using the random-effects model of Han and Eskin, implemented in *METASOFT*.[59] This model separates hypothesis testing from the estimation of the effect size, which allows the test to better model the between-study heterogeneity that is typically encountered in a trans-ethnic meta-analysis. Prior to using *METASOFT*, study-specific results were filtered as described above, which included removing genetic variants with eMAC < 10, and applying genomic control by multiplying each variant's standard error by the inverse of the square root of the lambda for cohorts with λ > 1. Genetic variants for which the total *Effective N* was less than half of the maximum *Effective N* were removed from the meta-analyses results.

7.4.6 Conditional analyses

Conditional analyses were performed using the *-condition-on* option implemented in *SNPTEST* to determine the number of independent signals at the *APOE* locus. We performed this analysis in the cohorts that were analysed using *SNPTEST* and for which both the ApoE ϵ4 and ApoE ϵ2 variant showed a significant association in the unadjusted analysis (*i.e.*, CEPH and LLS (combined with GEHA Dutch)). In both cohorts, the association of ApoE ϵ2 remained significant (p < 0.05) after adjustment for ApoE ϵ4, indicating an independent effect.

7.4.7 Gene-level association analysis

MetaXcan was used to identify genetically predicted tissue-specific expression associations with longevity using the results from the 90^{th} and 99^{th} percentile cases versus all controls meta-analyses.[24] GTEx version 7 tissue models of genetically predicted expression were used. To maximize the number of genetic variants that *MetaXcan* could match with tissue models, the *MetaXcan* SNP annotation file (gtex_v7_hapmapceu_dbsnp150_snp_annot.txt) was used to map variants from the GWA study results file to rsIDs by chromosome, position, and alleles. To control for the false discovery rate when testing multiple genes across multiple tissues, the Storey q-value was applied and a q-value < 0.05 was considered significant.[60] Colocalization of the tissue-specific eQTL results from the GTEx project and our longevity meta-analyses results was performed using the *coloc.abf* function implemented in the R-package *coloc*.[61]

7.4.8 Genetic correlation analysis

To estimate the genetic correlation between the different phenotypes used in this study, we used LD score regression.[23] The genetic correlation between the results from the 90^{th} and 99^{th} percentile cases versus all controls meta-analyses and 246 diseases and traits were estimated using the LD Hub web portal (`http://ldsc.broadinstitute.org/ldhub/`).[25] Since LD score regression is currently only possible with data from individuals of European ancestry, we used our meta-analyses results based on the cohorts from populations of European descent only.

7.4.9 Power calculation

The power calculations for the validation in the UK Biobank and for the replication of previously identified loci associated with human lifespan were performed using the Genetic Association Study Power Calculator (available here) using an additive disease model and a disease prevalence of 0.1 (90^{th} percentile) or 0.01 (99^{th} percentile).

7.4.10 Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

7.5 Data availability

The full meta-analyses summary statistics are available for download at this address, through GRASP website (accessible here), and through the NHGRI-EBI GWAS Catalog website (here). All other data that supports the findings of this study are available from the corresponding authors upon request.

7.6 Acknowledgements

A full list of acknowledgements, including support for each of the participating cohorts, is provided in Supplementary Note 1.

7.7 Full author list and affiliations

Joris Deelen, [1,2,78] Daniel S. Evans,[3,78] Dan E. Arking,[4] Niccolo' Tesi,[5,6,7] Marianne Nygaard,[8] Xiaomin Liu,[9,10] Mary K. Wojczynski, [18] Mary L. Biggs,[12,13] Ashley van der Spek,[14] Gil Atzmon,[15,16] Erin B. Ware,[17] Chloè Sarnowski,[18] Albert V. Smith,[19,20] Ilkka Seppala,[21] Heather J. Cordell,[22] Janina Dose,[23]

Najaf Amin,[14] Alice M. Arnold,[12] Kristin L. Ayers,[24] Nir Barzilai,[16] Elizabeth J. Becker,[25] Marian Beekman,[2] Hèlène Blanchè,[26] Kaare Christensen,[8,27,28] Lene Christiansen,[8,29] Joanna C. Collerton,[30] Sarah Cubaynes,[31] Steven R. Cummings,[3] Karen Davies,[32] Birgit Debrabant,[33] Jean-Francois Deleuze,[26,34] Rachel Duncan,[30,35] Jessica D. Faul,[17] Claudio Franceschi,[36,37] Pilar Galan,[38] Vilmundur Gudnason,[20,39] Tamara B. Harris,[40] Martijn Huisman,[41,42] Mikko A. Hurme,[43] Carol Jagger,[30,35] Iris Jansen,[5,44] Marja Jylha,[45] Mika Kahonen,[46] David Karasik,[47,48] Sharon L.R. Kardia,[49] Andrew Kingston,[30,35] Thomas B.L. Kirkwood,[35] Lenore J. Launer,[40] Terho Lehtimaki,[21] Wolfgang Lieb,[50] Leo-Pekka Lyytikainen,[21] Carmen Martin-Ruiz,[32] Junxia Min,[51] Almut Nebel,[23] Anne B. Newman,[52] Chao Nie,[9] Ellen A. Nohr,[53] Eric S. Orwoll,[54] Thomas T. Perls,[55] Michael A. Province,[11] Bruce M. Psaty,[13,56,57,58] Olli T. Raitakari,[59,60] Marcel J.T. Reinders,[7] Jean-Marie Robine,[31,61] Jerome I. Rotter,[62,63] Paola Sebastiani,[18] Jennifer Smith,[17,49] Thorkild I.A. Sorensen,[64,65] Kent D. Taylor,[62,66] André G. Uitterlinden,[14,67] Wiesje van der Flier,[5,41] Sven J. van der Lee,[5,6] Cornelia M. van Duijn,[14,68] Diana van Heemst,[69] James W. Vaupel,[70] David Weir,[17] Kenny Ye Yi Zeng,[72,73] Wanlin Zheng,[3] Henne Holstege,[5,6,7] Douglas P. Kiel,[48,74,75] Kathryn L. Lunetta,[18] P. Eline Slagboom,[2,78] and Joanne M. Murabito [76,77,78]

[1] Max Planck Institute for Biology of Ageing, 50866 Cologne, Germany.

[2] Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, 2300 RC Leiden, The Netherlands.

[3] California Pacific Medical Center Research Institute, San Francisco, CA 94158, USA.

[4] McKusick-Nathans Institute of Genetic Medicine, Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA.

[5] Alzheimer Center Amsterdam, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, 1007 MB Amsterdam, The Netherlands.

[6] Department of Clinical Genetics, Amsterdam UMC, 1007 MB Amsterdam, The Netherlands.

[7] Delft Bioinformatics Lab, Delft University of Technology, 2600 GA Delft, The Netherlands. [8] The Danish Aging Research Center, Department of Public Health, University of Southern Denmark, 5000 Odense C, Denmark.

[9] BGI-Shenzhen, Shenzhen 518083, China. [10] China National Genebank, BGI- Shenzhen, Shenzhen 518120, China.

[11] Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, Saint Louis, MO 63110, USA.

[12] Department of Biostatistics, Univer-

sity of Washington, Seattle, WA 98115, USA.
[13] Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA 98101, USA.
[14] Department of Epidemiology, Erasmus MC, 3000 CA Rotterdam, The Netherlands.
[15] Department of Biology, Faculty of Natural Science, University of Haifa, Haifa 3498838, Israel.
[16] Departments of Medicine and Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA.
[17] Institute for Social Research, Survey Research Center, University of Michigan, Ann Arbor, MI 48104, USA.
[18] Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA.
[19] School of Public Health, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA.
[20] Icelandic Heart Association, 201 Kópavogur, Iceland.
[21] Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center—Tampere, Faculty of Medicine and Health Technology, Tampere University, 33520 Tampere, Finland.
[22] Institute of Genetic Medicine, Newcastle University, Newcastle upon Tyne NE1 3BZ, UK.
[23] Institute of Clinical Molecular Biology, Kiel University, 24105 Kiel, Germany.
[24] Sema4, a Mount Sinai venture, Stamford, CT 06902, USA.
[25] Bioinformatics Program, Boston University, Boston, MA 02118, USA.
[26] Fondation Jean Dausset-CEPH, 75010 Paris, France.
[27] Clinical Biochemistry and Pharmacology, Odense University Hospital, 5000 Odense C, Denmark.
[28] Department of Clinical Genetics, Odense University Hospital, 5000 Odense C, Denmark.
[29] Department of Clinical Immunology, Copenhagen University Hospital, Rigshospitalet, 2100 Copenhagen, Denmark.
[30] Institute of Health and Society, Newcastle University, Newcastle upon Tyne NE4 5PL, UK.
[31] MMDN, Univ. Montpellier, EPHE, Unité Inserm 1198, PSL Research University, 34095 Montpellier, France.
[32] Institute of Neuroscience, Newcastle University, Newcastle upon Tyne NE4 5PL, UK.
[33] Department of Public Health, University of Southern Denmark, 5000 Odense C, Denmark.
[34] Centre National de Recherche en Gènomique Humaine, CEA-Institut de Biologie Francois Jacob, 91000 Evry, France.
[35] Newcastle University Institute for Ageing, Newcastle University, Newcastle upon Tyne NE4 5PL, UK.
[36] Department of Applied Mathematics and Centre of Bioinformatics, Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod 603022, Russia.

[37] IRCCS Institute of Neurological Sciences of Bologna (ISNB), 40124 Bologna, Italy.
[38] EREN, UMR U1153 Inserm/U1125 Inra/Cnam/Paris 13, Universitè Paris 13, CRESS, 93017 Bobigny, France.
[39] Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland.
[40] Laboratory of Epidemiology and Population Sciences, National Institute on Aging, NIH, Bethesda, MD 20892, USA.
[41] Department of Epidemiology and Biostatistics, Vrije Universiteit Amsterdam, Amsterdam UMC, 1007 MB Amsterdam, The Netherlands.
[42] Amsterdam Public Health Research Institute, 1007 MB Amsterdam, The Netherlands.
[43] Department of Microbiology and Immunology, Faculty of Medicine and Health Technology, Tampere University, 33014 Tampere, Finland.
[44] Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Vrije Universiteit Amsterdam, 1081 HV Amsterdam, The Netherlands.
[45] Faculty of Social Sciences (Health Sciences) and Gerontology Research Center (GEREC), Tampere University, 33104 Tampere, Finland.
[46] Department of Clinical Physiology, Tampere University Hospital and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, 33521 Tampere, Finland.
[47] Azrieli Faculty of Medicine, Bar Ilan University, Safed 13010, Israel.
[48] Hinda and Arthur Marcus Institute for Aging Research, Hebrew SeniorLife, Boston, MA 02131, USA.
[49] School of Public Health, Epidemiology, University of Michigan, Ann Arbor, MI 48109, USA.
[50] Institute of Epidemiology and Biobank PopGen, Kiel University, 24105 Kiel, Germany.
[51] Institute of Translational Medicine, School of Medicine, Zhejiang University, Hangzhou 311058, China.
[52] Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh,PA 15261, USA.
[53] Research Unit of Gynecology and Obstetrics, Department of Clinical Research, University of Southern Denmark, 5000 Odense C, Denmark.
[54] Bone and Mineral Unit, Oregon Health Sciences University, Portland, OR 97239, USA.
[55] Department of Medicine, Geriatrics Section, Boston Medical Center, Boston University School of Medicine, Boston, MA 02118, USA.
[56] Department of Epidemiology, University of Washington, Seattle, WA 98101, USA.
[57] Department of Health Services, University of Washington, Seattle, WA 98101, USA.
[58] Kaiser Permanente Washington Health Research Institute, Seattle, WA 98101, USA.
[59] Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, 20521 Turku, Finland.

[60] Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, 20014 Turku, Finland.
[61] CERMES3, UMR CNRS 8211-Unité Inserm 988-EHESS-Université Paris Descartes, 94801 Paris, France.
[62] Institute for Translational Genomics and Population Sciences, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA 90502, USA.
[63] Division of Genetic Outcomes, Department of Pediatrics, Harbor-UCLA Medical Center, Torrance, CA 90502, USA.
[64] Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, and Department of Public Health, Section of Epidemiology, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen N, Denmark.
[65] MRC Integrative Epidemiology Unit, Bristol University, BS8 2BN Bristol, UK.
[66] Department of Pediatrics, Harbor-UCLA Medical Center, Torrance, CA 90502, USA.
[67] Department of Internal Medicine, Erasmus MC, 3000 CA Rotterdam, The Netherlands.
[68] Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK.
[69] Department of Gerontology and Geriatrics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands.
[70] Max Planck Institute for Demographic Research, 18057 Rostock, Germany.
[71] Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA.
[72] Center for Healthy Aging and Development Studies, National School of Development and Raissun Institute for Advanced Studies, Peking University, 100871 Beijing, China.
[73] Center for the Study of Aging and Human Development and Geriatrics Division, Medical School of Duke University, Durham, NC 27710, USA.
[74] Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA 02215, USA.
[75] Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA.
[76] NHLBI's and Boston University's Framingham Heart Study, Framingham, MA 01702, USA.
[77] Section of General Internal Medicine, Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA.
[78] These authors contributed equally: Joris Deelen, Daniel S. Evans, P. Eline Slagboom, Joanne M. Murabito.

7.8 Supplementary Figures

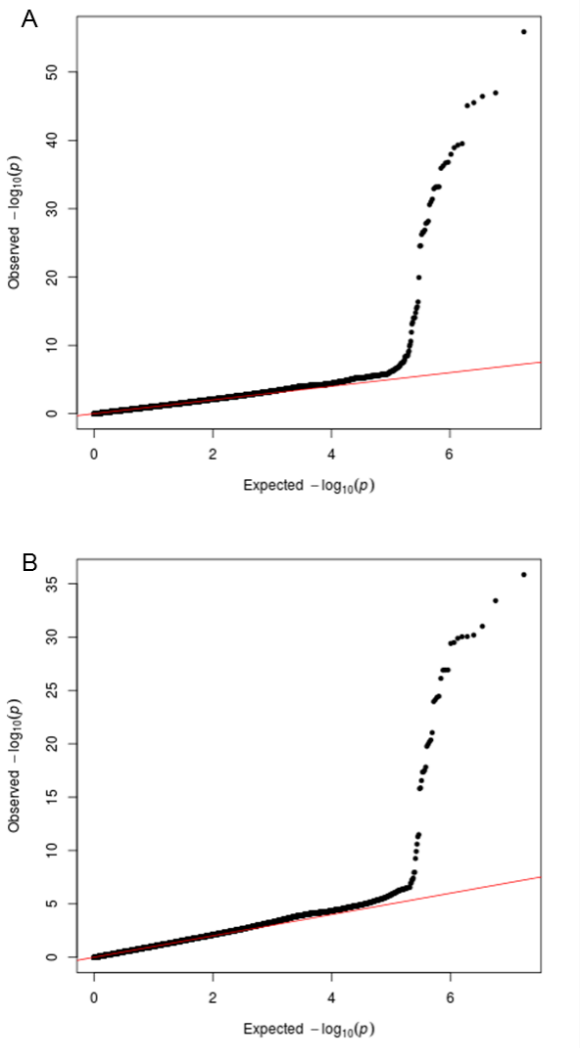


Figure 7.4: **Quantile-quantile plots for the European genome-wide association meta-analyses.** Quantile-quantile plots of the expected versus (unadjusted) observed $-\log_{10}$ *P*-values for the European genome-wide association meta-analyses of the 90^{th} percentile cases versus all controls (λ = 1.036, **a**) and 99^{th} percentile cases versus all controls (λ = 1.036, **b**).

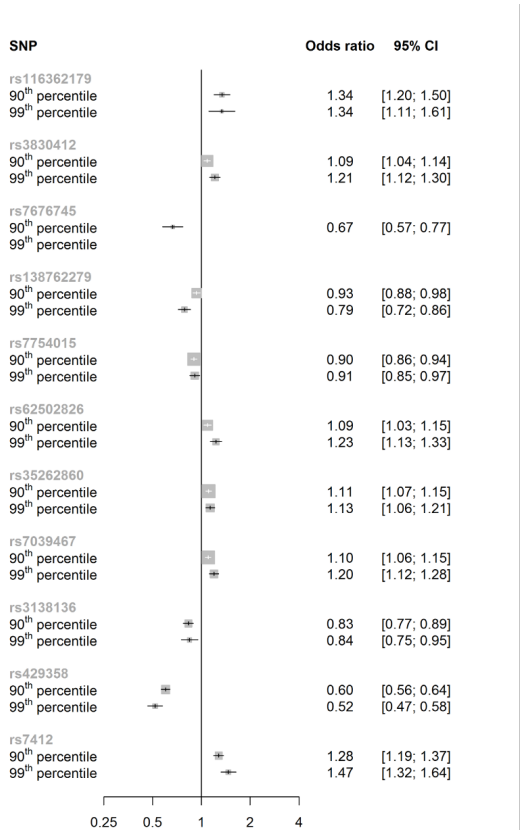


Figure 7.5: **Results of the suggestive significant genetic variants from the European genome-wide association meta-analyses.** Forest plot for the suggestive significant genetic variants from the European genome-wide association meta-analyses of the 90^{th} and 99^{th} percentile versus all controls. We had insufficient studies with data for *rs7676745* in the 99^{th} percentile versus all controls meta-analysis to reliable analyse this genetic variant due to its relatively low minor allele frequency.

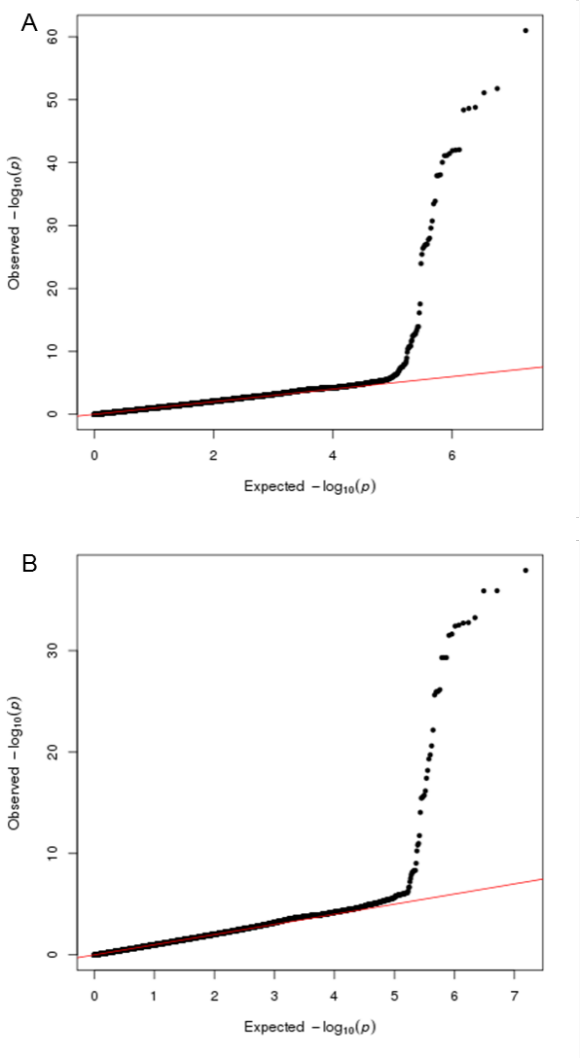


Figure 7.6: **Quantile-quantile plots for the trans-ethnic genome-wide association meta-analyses.** Quantile-quantile plots of the expected versus (unadjusted) observed $-\log_{10}$ *P*-values for the trans-ethnic genome-wide association meta-analysis of the 90^{th} percentile cases versus all controls (λ = 0.97, **a**) and 99^{th} percentile cases versus all controls (λ = 0.93, **b**).

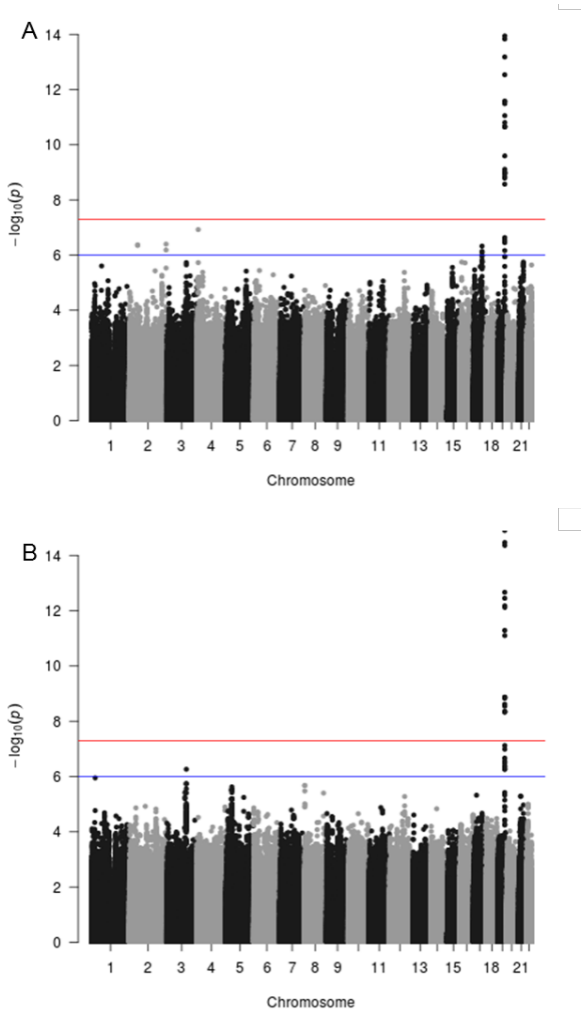


Figure 7.7: **Results for the European genome-wide association meta-analyses using different control definitions.** Manhattan plot presenting the $-\log_{10}$ *P*-values from the European genome-wide association meta-analysis of the 90^{th} percentile cases versus all controls (**a**) or dead controls only (**b**). The red line indicates the threshold for genome-wide significance ($P \leq 5\text{x}10^{-8}$), while the blue line indicates the threshold for genetic variants that showed a suggestive significant association ($P \leq 1\text{x}10^{-6}$). For representation purposes, the maximum of the Y-axis was set to 14.

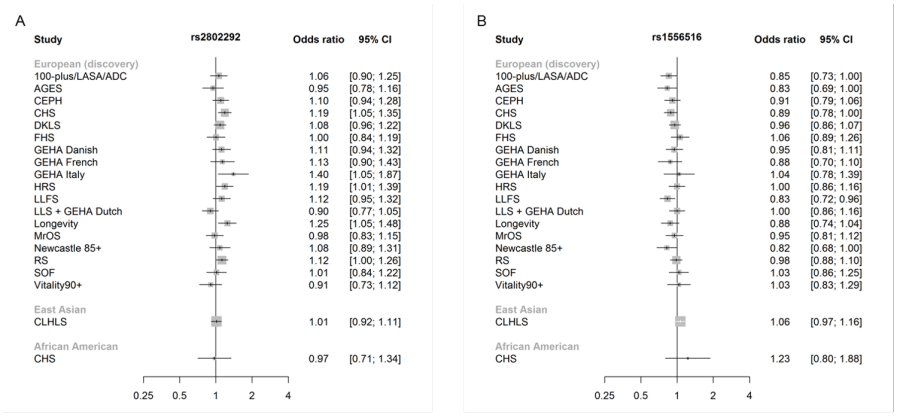


Figure 7.8: **Study-specific results for the genetic variants in FOXO3 and CDKN2A/B.** Forest plots for *rs2802292* (**a**) and *rs1556516* (**b**) based on the results from the 90^{th} percentile versus all controls analysis. The size of the boxes represents the sample size of the cohort.

7.9 Supplementary Tables

Supplementary Tables and Supplementary Information can be accessed by scanning the following code or accessing the journal's website at this address.

References

[1] J. Oeppen. "DEMOGRAPHY: Enhanced: Broken Limits to Life Expectancy". In: *Science* 296.5570 (May 2002), pp. 1029–1031. ISSN: 00368075, 10959203. DOI: 10.1126/science.1069675.

[2] Niels van den Berg et al. "Longevity defined as top 10% survivors and beyond is transmitted as a quantitative genetic trait". In: *Nature Communications* 10.1 (Dec. 2019), p. 35. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07925-0.

[3] Joanna Kaplanis et al. "Quantitative analysis of population-scale family trees with millions of relatives". In: *Science* 360.6385 (Apr. 13, 2018), pp. 171–175. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aam9309.

[4] J. Graham Ruby et al. "Estimates of the Heritability of Human Longevity Are Substantially Inflated due to Assortative Mating". In: *Genetics* 210.3 (Nov. 2018), pp. 1109–1124. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.118.301613.

[5] Paola Sebastiani et al. "Increasing Sibling Relative Risk of Survival to Older and Older Ages and the Importance of Precise Definitions of "Aging," "Life Span," and "Longevity"". In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 71.3 (Mar. 2016), pp. 340–346. ISSN: 1079-5006, 1758-535X. DOI: 10.1093/gerona/glv020.

[6] Linda Broer et al. "GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy". In: *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 70.1 (Jan. 2015), pp. 110–118. ISSN: 1758-535X. DOI: 10.1093/gerona/glu166.

[7] Joris Deelen et al. "Genome-wide association meta-analysis of human longevity identifies a novel locus conferring survival beyond 90 years of age". In: *Human Molecular Genetics* 23.16 (Aug. 2014), pp. 4420–4432. ISSN: 1460-2083. DOI: 10.1093/hmg/ddu139.

[8] Yi Zeng et al. "Novel loci and pathways significantly associated with longevity". In: *Scientific Reports* 6.1 (Aug. 2016). ISSN: 2045-2322. DOI: 10.1038/srep21243.

[9] Peter K. Joshi et al. "Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity". In: *Nature Communications* 8.1 (Dec. 2017). ISSN: 2041-1723. DOI: 10.1038/s41467-017-00934-5.

[10] Aaron F. McDaid et al. "Bayesian association scan reveals loci associated with human lifespan and linked biomarkers". In: *Nature Communications* 8 (July 2017), p. 15842. ISSN: 2041-1723. DOI: 10.1038/ncomms15842.

[11] Luke C. Pilling et al. "Human longevity: 25 genetic loci associated in 389,166 UK biobank participants". In: *Aging* 9.12 (Dec. 2017), pp. 2504–2520. ISSN: 1945-4589. DOI: 10.18632/aging.101334.

[12] Paola Sebastiani et al. "Four Genome-Wide Association Studies Identify New Extreme Longevity Variants". In: *The Journals of Gerontology: Series A* 72.11 (Oct. 2017), pp. 1453–1464. ISSN: 1079-5006, 1758-535X. DOI: 10.1093/gerona/glx027.

[13] Paul RHJ Timmers et al. "Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances". In: *eLife* 8 (Jan. 2019). ISSN: 2050-084X. DOI: 10.7554/eLife.39856.

[14] Friederike Flachsbart et al. "Immunochip analysis identifies association of the RAD50/IL13 region with human longevity". In: *Aging Cell* 15.3 (2016), pp. 585–588. ISSN: 1474-9726. DOI: 10.1111/acel.12471.

[15] Anne B. Newman et al. "A meta-analysis of four genome-wide association studies of survival to age 90 years or older: the Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium". In: *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 65.5 (May 2010), pp. 478–487. ISSN: 1758-535X. DOI: 10.1093/gerona/glq028.

[16] Almut Nebel et al. "A genome-wide association study confirms APOE as the major gene influencing survival in long-lived individuals". In: *Mechanisms of Ageing and Development* 132.6 (June 2011), pp. 324–330. ISSN: 00476374. DOI: 10.1016/j.mad.2011.06.008.

[17] Joris Deelen et al. "Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited: GWAS for familial longevity; APOE locus revisited". In: *Aging Cell* 10.4 (Aug. 2011), pp. 686–698. ISSN: 14749718. DOI: 10.1111/j.1474-9726.2011.00705.x.

[18] Yi Zeng et al. "Sex Differences in Genetic Associations With Longevity". In: *JAMA network open* 1.4 (2018), e181670. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2018.1670.

[19] Anatoliy I Yashin et al. "Genetics of Human Longevity From Incomplete Data: New Findings From the Long Life Family Study". In: *The Journals of Gerontology: Series A* 73.11 (Oct. 8, 2018), pp. 1472–1481. ISSN: 1079-5006, 1758-535X. DOI: 10.1093/gerona/gly057.

[20] Linda Partridge, Joris Deelen, and P. Eline Slagboom. "Facing up to the global challenges of ageing". In: *Nature* 561.7721 (Sept. 2018), pp. 45–56. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0457-8.

[21] M. L. Miller Bell F. C. "Life Tables for the United States Social Security Area 1900–2100." In: *SSA Pub. No.* (2005), pp. 11–11536.

[22] Marianne Nygaard et al. "Birth cohort differences in the prevalence of longevity-associated variants in APOE and FOXO3A in Danish long-lived individuals". In: *Experimental Gerontology* 57 (Sept. 2014), pp. 41–46. ISSN: 1873-6815. DOI: 10.1016/j.exger.2014.04.018.

[23] Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature Genetics* 47.3 (Mar. 2015), pp. 291–295. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.3211.

[24] Alvaro N. Barbeira et al. "Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics". In: *Nature Communications* 9.1 (2018), p. 1825. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03621-1.

[25] Jie Zheng et al. "LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis". eng. In: *Bioinformatics (Oxford, England)* 33.2 (2017), pp. 272–279. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw613.

[26] F. Schächter et al. "Genetic associations with human longevity at the APOE and ACE loci". In: *Nature Genetics* 6.1 (Jan. 1994), pp. 29–32. ISSN: 1061-4036. DOI: 10.1038/ng0194-29.

[27] Paola Sebastiani et al. "APOE Alleles and Extreme Human Longevity". In: *The Journals of Gerontology: Series A* (July 2018). ISSN: 1079-5006, 1758-535X. DOI: 10.1093/gerona/gly174.

[28] Seungjin Ryu et al. "Genetic landscape of APOE in human longevity revealed by high-throughput sequencing". In: *Mechanisms of Ageing and Development* 155 (Apr. 2016), pp. 7–9. ISSN: 00476374. DOI: 10.1016/j.mad.2016.02.010.

[29] Nuria Garatachea et al. "ApoE gene and exceptional longevity: Insights from three independent cohorts". In: *Experimental Gerontology* 53 (May 2014), pp. 16–23. ISSN: 05315565. DOI: 10.1016/j.exger.2014.02.004.

[30] Guodong Liu et al. "APOE gene polymorphism in long-lived individuals from a central China population". In: *Scientific Reports* 7.1 (Dec. 2017). ISSN: 2045-2322. DOI: 10.1038/s41598-017-03227-5.

[31] R. W. Mahley and S. C. Rall. "Apolipoprotein E: far more than a lipid transport protein". In: *Annual Review of Genomics and Human Genetics* 1 (2000), pp. 507–537. ISSN: 1527-8204. DOI: 10.1146/annurev.genom.1.1.507.

[32] Yadong Huang et al. "Apolipoprotein E: diversity of cellular origins, structural and biophysical properties, and effects in Alzheimer's disease". In: *Journal of molecular neuroscience: MN* 23.3 (2004), pp. 189–204. ISSN: 0895-8696. DOI: 10.1385/JMN:23:3:189.

[33] James R. Staley et al. "PhenoScanner: a database of human genotype-phenotype associations". In: *Bioinformatics (Oxford, England)* 32.20 (2016), pp. 3207–3209. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw373.

[34] Daniel M. Rosenbaum, Søren G. F. Rasmussen, and Brian K. Kobilka. "The structure and function of G-protein-coupled receptors". In: *Nature* 459.7245 (May 21, 2009), pp. 356–363. ISSN: 1476-4687. DOI: 10.1038/nature08144.

[35] Dan-Dan Dong, Hui Zhou, and Gao Li. "GPR78 promotes lung cancer cell migration and metastasis by activation of Gαq-Rho GTPase pathway". In: *BMB reports* 49.11 (Nov. 2016), pp. 623–628. ISSN: 1976-670X. DOI: 10.5483/bmbrep.2016.49.11.133.

[36] Valentina Grossi et al. "The longevity SNP rs2802292 uncovered: HSF1 activates stress-dependent expression of FOXO3 through an intronic enhancer". In: *Nucleic Acids Research* 46.11 (2018), pp. 5587–5600. ISSN: 1362-4962. DOI: 10.1093/nar/gky331.

[37] Friederike Flachsbart et al. "Identification and characterization of two functional variants in the human longevity gene FOXO3". In: *Nature Communications* 8.1 (2017), p. 2063. ISSN: 2041-1723. DOI: 10.1038/s41467-017-02183-y.

[38] Luke C. Pilling et al. "Human longevity is influenced by many genetic variants: evidence from 75,000 UK Biobank participants". In: *Aging* 8.3 (Mar. 2016), pp. 547–560. ISSN: 1945-4589. DOI: 10.18632/aging.100930.

[39] William R. Jeck, Alex P. Siebold, and Norman E. Sharpless. "Review: a meta-analysis of GWAS and age-associated diseases". In: *Aging Cell* 11.5 (Oct. 2012), pp. 727–731. ISSN: 1474-9726. DOI: 10.1111/j.1474-9726.2012.00871.x.

[40] Heribert Schunkert et al. "Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease". In: *Nature Genetics* 43.4 (Mar. 6, 2011), pp. 333–338. ISSN: 1546-1718. DOI: 10.1038/ng.784.

[41] Darren J. Baker, Fang Jin, and Jan M. van Deursen. "The yin and yang of the Cdkn2a locus in senescence and aging". In: *Cell Cycle (Georgetown, Tex.)* 7.18 (Sept. 15, 2008), pp. 2795–2802. ISSN: 1551-4005. DOI: 10.4161/cc.7.18.6687.

[42] Carlos López-Otín et al. "The hallmarks of aging". In: *Cell* 153.6 (June 6, 2013), pp. 1194–1217. ISSN: 1097-4172. DOI: 10.1016/j.cell.2013.05.039.

[43] Anne B. Newman et al. "Health and function of participants in the Long Life Family Study: A comparison with other cohorts". In: *Aging* 3.1 (Jan. 2011), pp. 63–76. ISSN: 1945-4589. DOI: 10.18632/aging.100242.

[44] Rudi G.J. Westendorp et al. "Nonagenarian Siblings and Their Offspring Display Lower Risk of Mortality and Morbidity than Sporadic Nonagenarians: The Leiden Longevity Study: MORTALITY RISK AND DISEASE PREVALENCE IN FAMILIAL LONGEVITY". In: *Journal of the American Geriatrics Society* 57.9 (Sept. 2009), pp. 1634–1637. ISSN: 00028614, 15325415. DOI: 10.1111/j.1532-5415.2009.02381.x.

[45] Lori Mosca, Elizabeth Barrett-Connor, and Nanette Kass Wenger. "Sex/gender differences in cardiovascular disease prevention: what a difference a decade makes". In: *Circulation* 124.19 (Nov. 8, 2011), pp. 2145–2154. ISSN: 1524-4539. DOI: 10.1161/CIRCULATIONAHA.110.968792.

[46] E. A. Gale and K. M. Gillespie. "Diabetes and gender". In: *Diabetologia* 44.1 (Jan. 2001), pp. 3–15. ISSN: 0012-186X. DOI: 10.1007/s001250051573.

[47] Marie Ng et al. "Smoking prevalence and cigarette consumption in 187 countries, 1980-2012". In: *JAMA* 311.2 (Jan. 8, 2014), pp. 183–192. ISSN: 1538-3598. DOI: 10.1001/jama.2013.284692.

[48] Stacy L. Andersen et al. "Health span approximates life span among many supercentenarians: compression of morbidity at the approximate limit of life span". In: *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 67.4 (Apr. 2012), pp. 395–405. ISSN: 1758-535X. DOI: 10.1093/gerona/glr223.

[49] Paola Sebastiani et al. "Limitations and risks of meta-analyses of longevity studies". In: *Mechanisms of Ageing and Development* 165 (Pt B 2017), pp. 139–146. ISSN: 1872-6216. DOI: 10.1016/j.mad.2017.01.008.

[50] CARDIoGRAMplusC4D Consortium et al. "Large-scale association analysis identifies new risk loci for coronary artery disease". In: *Nature Genetics* 45.1 (Jan. 2013), pp. 25–33. ISSN: 1546-1718. DOI: 10.1038/ng.2480.

[51] DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium et al. “Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility”. In: *Nature Genetics* 46.3 (Mar. 2014), pp. 234–244. ISSN: 1546-1718. DOI: 10.1038/ng.2897.

[52] Aleksandr Zenin et al. “Identification of 12 genetic loci associated with human healthspan”. In: *Communications Biology* 2 (2019), p. 41. ISSN: 2399-3642. DOI: 10.1038/s42003-019-0290-0.

[53] Jose Lara et al. “A proposed panel of biomarkers of healthy ageing”. In: *BMC medicine* 13 (Sept. 15, 2015), p. 222. ISSN: 1741-7015. DOI: 10.1186/s12916-015-0470-9.

[54] Genevieve L. Wojcik et al. “Genetic analyses of diverse populations improves discovery for complex traits”. In: *Nature* 570.7762 (2019), pp. 514–518. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1310-4.

[55] {and} Max Planck Institute for Demographic Research (Germany). University of California Berkeley (USA). “Human Mortality Database.” In: ().

[56] Alexander Kulminski, Irina Culminskaya, and Anatoli I Yashin. “Letter to the editor: Standardization of genetic association studies, pros and cons, reaffirmed”. In: *AGE* 36.2 (Apr. 2014), pp. 945–947. ISSN: 0161-9152, 1574-4647. DOI: 10.1007/s11357-013-9602-3.

[57] Thomas W. Winkler et al. “Quality control and conduct of genome-wide association meta-analyses”. In: *Nature Protocols* 9.5 (May 2014), pp. 1192–1212. ISSN: 1750-2799. DOI: 10.1038/nprot.2014.071.

[58] Cristen J. Willer, Yun Li, and Gonçalo R. Abecasis. “METAL: fast and efficient meta-analysis of genomewide association scans”. In: *Bioinformatics (Oxford, England)* 26.17 (Sept. 1, 2010), pp. 2190–2191. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btq340.

[59] Buhm Han and Eleazar Eskin. “Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies”. In: *American Journal of Human Genetics* 88.5 (May 13, 2011), pp. 586–598. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2011.04.014.

[60] John D. Storey and Robert Tibshirani. “Statistical significance for genomewide studies”. In: *Proceedings of the National Academy of Sciences of the United States of America* 100.16 (Aug. 5, 2003), pp. 9440–9445. ISSN: 0027-8424. DOI: 10.1073/pnas.1530509100.

[61] Claudia Giambartolomei et al. “Bayesian test for colocalisation between pairs of genetic association studies using summary statistics”. In: *PLoS genetics* 10.5 (May 2014), e1004383. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004383.

8. snpXplorer

snpXplorer: a web application to explore SNP-associations and annotate SNP-sets

Niccolo' Tesi, Sven J. van der Lee, Marc Hulsman, Henne Holstege and Marcel J.T. Reinders

This chapter was published in *Nucleic Acid Research*
https://doi.org/10.1093/nar/gkab410

Abstract

Genetic association studies are frequently used to study the genetic basis of numerous human phenotypes. However, the rapid interrogation of how well a certain genomic region associates across traits as well as the interpretation of genetic associations is often complex and requires the integration of multiple sources of annotation, which involves advanced bioinformatic skills. We developed *snpXplorer*, an easy-to-use web-server application for exploring Single Nucleotide Polymorphisms (SNP) association statistics and to functionally annotate sets of SNPs. *snpXplorer* can superimpose association statistics from multiple studies, and displays regional information including SNP associations, structural variations, recombination rates, eQTL, linkage disequilibrium patterns, genes and gene-expressions per tissue. By overlaying multiple GWAS studies, *snpXplorer* can be used to compare levels of association across different traits, which may help the interpretation of variant consequences. Given a list of SNPs, *snpXplorer* can also be used to perform variant-to-gene mapping and gene-set enrichment analysis to identify molecular pathways that are overrepresented in the list of input SNPs. *snpXplorer* is freely available at `https://snpxplorer.net`. Source code, documentation, example files and tutorial videos are available within the Help section of *snpXplorer* and at `https://github.com/TesiNicco/snpXplorer`.

8.1 Background

Genome-wide association studies (GWAS) and sequencing-based association studies are a powerful approach to investigate the genetic basis of complex human phenotypes and their heritability. Facilitated by the cost-effectiveness of both genotyping and sequencing methods and by established analysis guidelines, the number of genetic association studies has risen steeply in the last decade: as of February 2021, the GWAS-Catalog, a database of genetic association studies, contained 4,865 publications and 247,051 variant-trait associations. [1] To understand how genetic factors affect different traits, it is valuable to explore various annotations of genomic regions as well as how associations relate between different traits. But this requires combining diverse sources of annotation such as observed structural variations (SV), expression-quantitative-trait-loci (eQTL), or chromatin context. Moreover, a framework to quickly visualize and compare association statistics of specific genomic regions across multiple traits is missing, and may be beneficial to the community of researchers working on human genetics. In addition, the functional interpretation of the effects of genetic variants on a gene-, protein- or pathway-level is difficult as often genetic variants lie in non-coding regions of the genome. As a one- to one mapping between genetic variants and affected genes is not trivial in these circumstances, it might be wise to associate multiple genes with a variant. Hence, a profound knowledge of biological databases, bioinformatics tools, and programming skills is often required to interpret GWAS outcomes. Unfortunately, not everyone is equipped with these skills. To assist human geneticists, we have developed *snpXplorer*, a web-server application written in R that allows (*i*) the rapid exploration of any region in the genome with customizable genomic features, (*ii*) the superimposition of summary statistics from multiple genetic association studies, and (*iii*) the functional annotation and pathway enrichment analysis of SNP sets in an easy-to-use user interface.

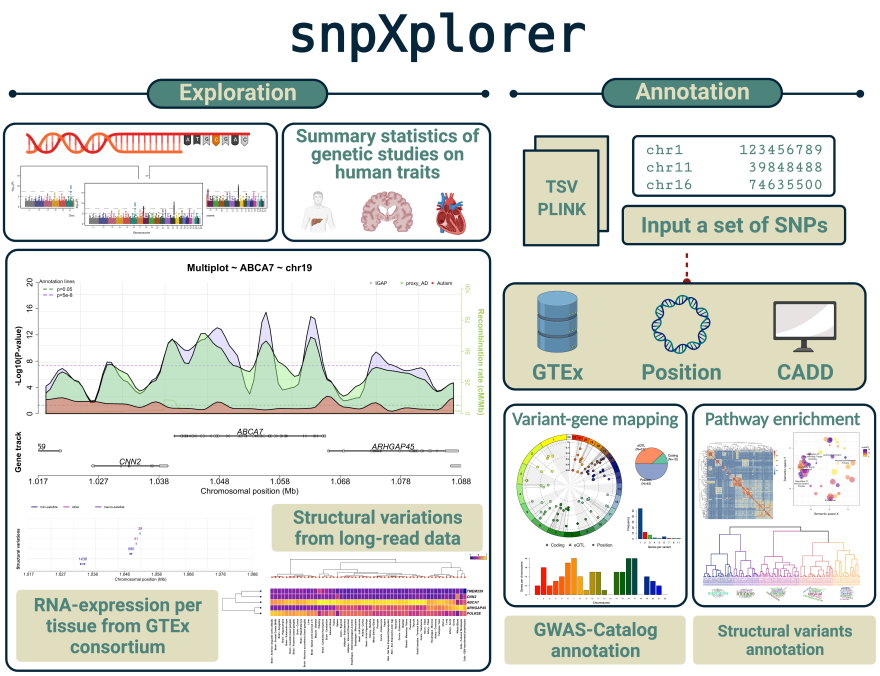


Figure 8.1: **snpXplorer graphical abstract.** The figure shows an overview of the exploration and annotation capabilities of *snpXplorer*.

8.2 Methods

8.2.1 Web server structure

snpXplorer is a web-server application based on the R package *shiny* that offers an exploration section and a functional annotation section. The exploration section represents the main interface (Figure 8.1) and provides an interactive exploration of a (set of) GWAS data sets. The functional annotation section takes as input any list of SNPs, runs a functional annotation and enrichment analysis in the background, and send the results by email.

8.2.2 Exploration section

First, input data must be chosen, which can either be one of the available summary statistics datasets and/or the user can upload their own association dataset. One of the main novelties in *snpXplorer* is the possibility to select multiple association datasets as inputs (including data uploaded by the user). These will be displayed on top of each other with different colours. The avail-

able summary statistics will be kept updated. As of February 2021, *snpXplorer* includes genome-wide summary statistics of 23 human traits classified in 5 disease categories: neurological traits (Alzheimer's disease, family history of Alzheimer's disease, autism, depression, and ventricular volume),[2, 3, 4, 5, 6] cardiovascular traits (coronary artery disease, systolic blood pressure, body-mass index and diabetes), [7, 8, 9, 10] immune-related traits (severe COVID infections, Lupus erythematosus, inflammation biomarkers and asthma),[11, 12, 13, 14] cancer-related traits (breast, lung, prostate cancers, myeloproliferative neoplasms and Lymphocytic leukaemia),[15, 16, 17, 18] and physiological traits (parental longevity, height, education, bone-density and vitamin D intake).[9, 19, 20, 21, 22] These summary statistics underwent a process of harmonization: we use the same reference genome (GRCh37, hg19) for all SNP positions, and in case a study was aligned to the GRCh38 (hg38), we translate the coordinates using the liftOver tool.[23] In addition, we only store chromosome, position and *p*-value information for each SNP-association. The user may upload own association statistics to display within *snpXplorer*: the file must have at least chromosome-, position-, and *p*-value columns, and the size should not exceed 600Mb. *snpXplorer* automatically recognizes the different columns, supports *PLINK* (v1.9+ and v2.0+) association files,[24] and we provide several example files in the Help section of the web-server. After selecting the input type, the user should set the preferred genome version. By default, GRCh37 is used, however, all available annotation sources are available also for GRCh38, and *snpXplorer* can translate genomic coordinates from one reference version to another. In order to browse the genome, the user can either input a specific genomic position, gene name, variant identifier, or select the scroll option, which allows to interactively browse the genome. The explorative visualisation consists of 3 separate panels showing (*i*) the SNP summary statistics of the selected input data (Figure 1A), (*ii*) the structural variants in the region of interest (Figure 1B), and (*iii*) the tissue-specific RNA-expression (Figure 1C). The first (and main) visualization panel shows the association statistics of the input data in the region of interest: genomic positions are shown on the x-axis and association significance (in $-\log_{10}$ scale) is reported on the y-axis. Both the x-axis and the y-axis can be interactively adjusted to extend or contract the genomic window to be displayed. Linkage disequilibrium (LD) patterns are optionally shown for the most significant variant in the region, the input variant, or a different variant of choice. The linkages are calculated using the genotypes of the individuals from the 1000Genome project, with the possibility to select the populations to include.[25] There are two ways to

visualise the data: by default, each variant-association is represented as a dot, with dot-sizes optionally reflecting *p*-values. Alternatively, associations can be shown as *p*-value profiles: to do so, (*i*) the selected region is divided in bins, (*ii*) a local maximum is found in each bin based on association *p*-value, and (*iii*) a polynomial regression model is fitted to the data, using the *p*-value of all local maximum points as dependent variable and their genomic position as predictors. Regression parameters, including the number of bins and the smoothing value, can be adjusted. Gene names from RefSeq (v98) are always adapted to the plotted region.[26] Finally, recombination rates from HapMap II, which give information about recombination frequency during meiosis, are optionally shown in the main plot interface.[27] The second panel shows structural variations (SV) in the region of interest. These are extracted from three studies that represent the state-of-the-art regarding the estimation of major structural variations across the genome using third-generation sequencing technologies (*i.e.* long read sequencing).[28, 29, 30] Structural variations are represented as segments: the size of the segment codes for the maximum difference in allele sizes of the SVs as observed in the selected studies. Depending on the different studies, structural variations are annotated as insertions, deletions, inversions, copy number alterations, duplications, mini-, micro- and macro-satellites, and mobile element insertions (Alu elements, LINE1 elements, and SVAs). The third panel shows tissue-specific RNA-expression (from the Genotype-Tissue-expression consortium, GTEx) of the genes displayed in the selected genomic window.[31] The expression of these genes across 54 human tissues is scaled and reported as a heatmap. Hierarchical clustering is applied on both the genes and the tissues, and the relative dendrograms are reported on the sides of the heatmap.

The side panel allows the user to interact with the exploration section. In order to guide the user through all the available inputs and options, help messages automatically appear upon hovering over items. The side panel reports (*i*) the top 10 variants with highest significance (together with the trait they belong to, in case multiple studies were selected), and (*ii*) the top eQTLs associations (by default, eQTLs in blood are shown, and this can be optionally changed), and cross-references including GeneCards, GWAS-catalog, and LD-hub.[1, 32, 33] Finally, download buttons allow to download a high-quality image of the different visualisation panels as well as the tables reporting the top SNP and eQTL associations, the SVs in the selected genomic window, and the LD table.

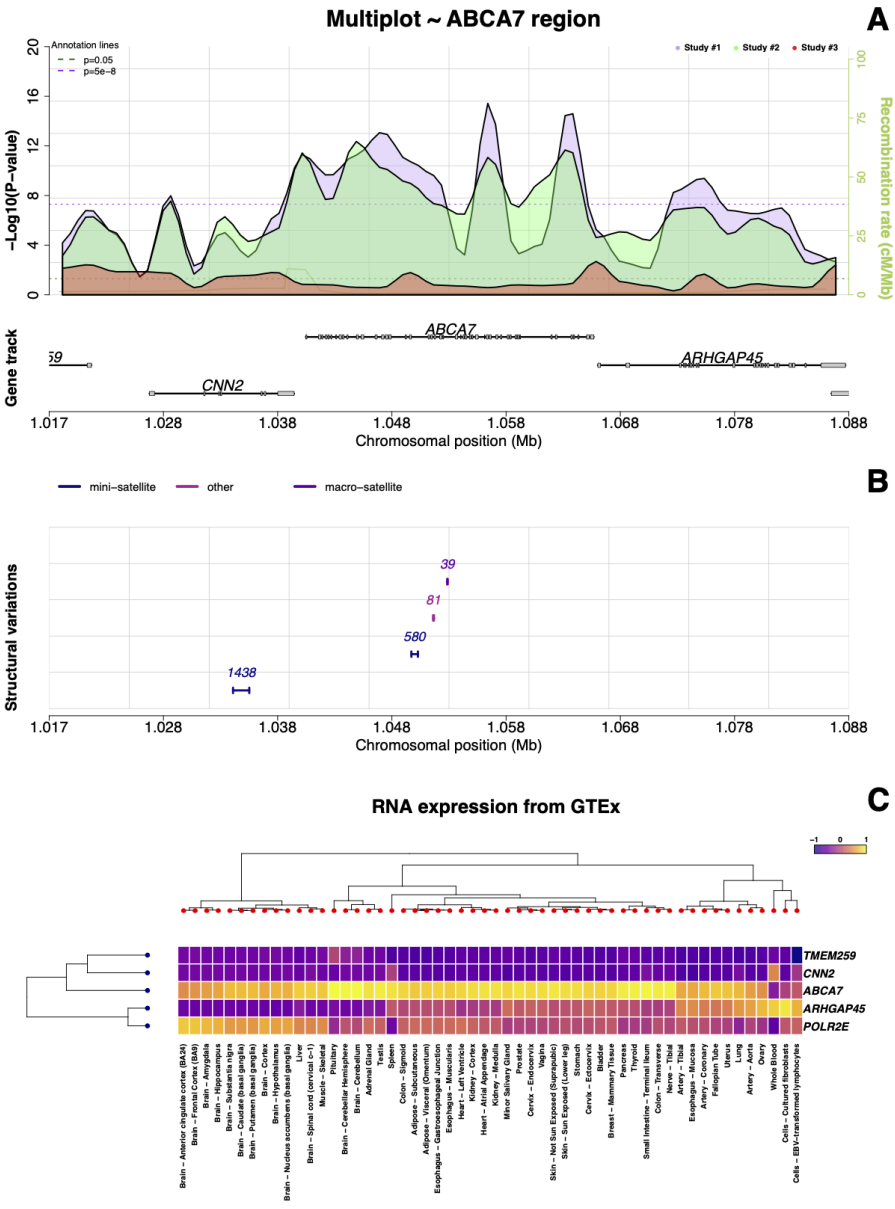


Figure 8.2: **snpXplorer exploration section. A.** First and main visualisation interface reporting summary statistics of multiple genetic studies as shown with *p*-value profiles. **B.** Structural variants within the region of interest are reported as segments and colored according to their type **C.** Tissue-specific RNA-expression (from Genotype-Tissue-Expression, GTEx) of the genes displayed in the region of interest.

8.2.3 Functional Annotation section

The functional annotation pipeline consists of a two-step procedure: firstly, genetic variants are linked to likely affected genes (variant-gene mapping); and, secondly, the likely affected genes are tested for pathway enrichment (gene-pathway mapping). In the variant-gene mapping, genetic variants are linked to the most likely affected gene(s) by (*i*) associating a variant to a gene when the variant is annotated to be coding by the Combined Annotation Dependent Depletion (CADD, v1.3), (*ii*) annotating a variant to genes based on found expression-quantitative-trait-loci (eQTL) from GTEx (v8, with possibility to choose the tissue(s) of interest), or (*iii*) mapping a variant to genes that are within distance d from the variant position, starting with $d \leq 50kb$, up to $d \geq 500kb$, increasing by 50kb until at least one match is found (from RefSeq v98).[26, 31, 34] Note that this procedure might map multiple genes to a single variant, depending on the effect and position of each variant. Then, we first report whether the input SNPs as well as their likely associated genes were previously associated with any trait in the GWAS-Catalog (traits are coded by their Experimental Factor Ontology (EFO) term). For this analysis, we downloaded all significant SNP-trait associations of all studies available in the GWAS-Catalog (v1.0.2, available at `https://www.ebi.ac.uk/gwas/docs/file-downloads`), which includes associations with $p < 9x10^{-6}$. Given a set of input SNPs associated with a set of genes, this analysis results in a set of traits (provided that the SNPs and/or the genes were previously associated with a trait). Hereto, we plot the number of SNPs in the list of uploaded SNPs that associate with the trait (expressed as a fraction). To correct for multiple genes being associated with a single variant, we estimate these fractions by sampling (500 iterations) one gene from the pool of genes associated with each variant, and averaging the resulting fractions across the sampling. Summary tables of the GWAS-Catalog analysis, including also EFO URI links for cross-referencing are provided as additional output. Next, we report on the structural variations that lie in the vicinity ($10kb$ upstream and downstream) of the input SNPs, and present information such as SV start and end position, SV type, maximum difference in allele size, and genes likely associated with the relative SNPs. Finally, we perform a gene-set enrichment analysis to find molecular pathways enriched within the set of genes associated with the input variants. Also, here we use the mentioned sampling technique to avoid a potential enrichment bias due to multiple genes being mapped to the same variant (this time the sampling is used to calculate p-values for each term). The gene-set enrichment analysis

is performed using the *Gost* function from the R package *gprofiler2*.[35] The user can specify several gene-set sources, such as Gene Ontology (release 2020-12-08),[36] KEGG (release 2020-12-14),[37] Reactome (release 2020-12-15),[38] and Wiki-pathways (release 2020-12-10).[39] The full table of the gene-set enrichment analysis comprising all tested terms and their relative sampling-based *p*-values is sent to the user. For each of the selected gene-set sources, the significant enriched terms are plotted (up to FDR<10%). In case the Gene Ontology is chosen as gene-set source, we additionally reduce the visual complexity of the enriched biological processes using (*i*) the *REVIGO* tool and (*ii*) a term-based clustering approach.[40] We do so because the interpretation of gene-set enrichment analyses is typically difficult due to the large number of terms. Clustering enriched terms then helps to get an overview, and thus eases the interpretation of the results. Briefly, *REVIGO* masks redundant terms based on a semantic similarity measure, and displays enrichment results in an embedded space via eigenvalue decomposition of the pairwise distance matrix. In addition to *REVIGO*, we developed a term-based clustering approach to remove redundancy between enriched terms. To do so, we first calculate a semantic similarity matrix between all enriched terms, and then apply hierarchical clustering on the obtained distance matrix. We estimate the optimal number of clusters using a dynamic cut tree algorithm and plot the most recurring words of the terms underlying each cluster using wordclouds. We use Lin as semantic distance measure for both *REVIGO* and our term-based clustering approach.[41] Figures representing *REVIGO* results, the semantic similarity heatmap (showing relationships between enriched terms), the hierarchical clustering dendrogram, and the wordclouds of each clusters, are generated. Finally, all tables describing *REVIGO* analysis and our term-based clustering approach (including all enriched terms and their clustering scheme) are produced and sent as additional output to the user for further manipulation. Note that the initial significant GO terms are not removed and also included in the reporting.

8.3 Results

8.3.1 Case Study

To illustrate the performances of *snpXplorer*, we explored the most recent set of common SNPs associated with late-onset Alzheimer's disease (AD, N=83 SNPs, Table S1).[42] Using this dataset as case study, we show the benefits of using *snpXplorer* in a typical scenario. Briefly, AD is the most prevalent type of dementia at old age, and is associated with a progressive loss of cognitive functions, ultimately leading to death. In its most common form (late-onset AD, with age at onset typically >65 years), the disease is estimated to be 60-80% heritable. With an attributable risk of ~30%, genetic variants in *APOE* gene represent the largest common genetic risk factor for AD. In addition to *APOE*, the genetic landscape of AD now counts 83 common variants that are associated with a slight modification of the risk of AD. Understanding the genes most likely involved in AD pathogenesis as well as the crucial biological pathways is warranted for the development of novel therapeutic strategies for AD patients. We retrieved the list of AD-associated genetic variants in Table 1 of the preprint from *Bellenguez et al, 2020*.[42] This study represent the largest GWAS on AD performed to date, and resulted in 42 novel SNPs reaching genome-wide evidence of association with AD. The exploration section of *snpXplorer* can be firstly used to inspect the association statistics of the novel SNP-associations in previous studies of the same trait (*i.e.* International Genomics of Alzheimer Project (IGAP) and family history of AD (proxy AD)). Specifically, a suggestive degree of association in these regions is expected to be found in earlier studies. As expected, suggestive association signals were already observed for the novel SNPs, increasing the likelihood that these novel SNPs are true associations (Figure 8.4).

After the first explorative analysis, we pasted the variant identifiers (rsIDs) in the annotation section of *snpXplorer*, specifying *rsid* as input type, Gene Ontology and Reactome as gene-sets for the enrichment analysis, and Blood as GTEx tissue for eQTL (*i.e.* the default value). The N=83 variants were linked to a total of 162 genes, with N=54 variants mapping to 1 gene, N=12 variants mapping to 2 genes, N=7 variants mapping to 3 genes, N=2 variants mapping to 4 genes, N=1 variant mapping to 5 genes, N=4 variants mapping to 4 genes, and N=1 variant mapping to 7, 8 and 11 genes (**??**). N=10 variants were found to be coding variants, N=31 variants were found to be eQTL, and N=42 variants were annotated based on their genomic position. These results are returned to the user in the form of a (human and machine-readable) table, but also in the form of a summary plot (Figure 8.3A and **??**).

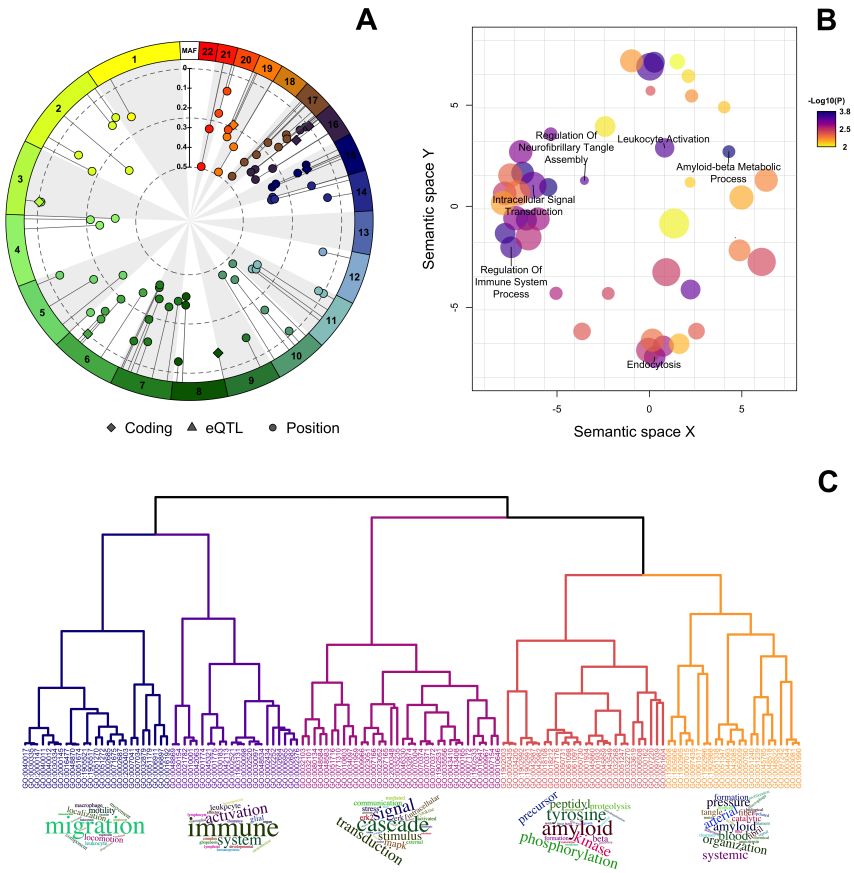


Figure 8.3: **Results of the functional annotation of N=83 variants associated with Alzheimer's disease (AD). A.** The circular summary figure shows the type of annotation of each genetic variant used as input (coding, eQTL or annotated by their positions) as well as each variant's minor allele frequency and chromosomal distribution. **B.** *REVIGO* plot, showing the remaining GO terms after removing redundancy based on a semantic similarity measure. The colour of each dot codes for the significance (the darker, the more significant), while the size of the dot codes for the number of similar terms removed from *REVIGO*. **C.** Results of our term-based clustering approach. We used Lin as semantic similarity measure to calculate similarity between all GO terms. We then used ward-d2 as clustering algorithm, and a dynamic cut tree algorithm to highlight clusters. Finally, for each cluster we generated wordclouds of the most frequent words describing each cluster.

These graphs not only inform the user about the effect of the SNPs of interest (for example, a direct consequence on the protein sequence in case of coding SNPs, or a regulatory effect in case of eQTLs or intergenic SNPs), but also suggest the presence of more complex regions: for example, **??**B indicates the number of genes associated with each SNP, which normally increases for complex, gene-dense regions such as *HLA*-region or *IGH*-region. In order to prioritize candidate genes, the authors of the original publication integrated (*i*) eQTLs and colocalization (eQTL coloc) analyses combined with expression transcriptome-wide association studies (eTWAS) in AD-relevant brain regions; (*ii*) splicing quantitative trait loci (sQTLs) and colocalization (sQTL coloc) analyses combined with splicing transcriptome-wide association studies (sTWAS) in AD-relevant brain regions; (*iii*) genetic-driven methylation as a biological mediator of genetic signals in blood (MetaMeth).[43] In order to compare the SNP-gene annotation of the original study with that of *snpXplorer*, we counted the total number of unique genes associated with the SNPs (*i*) in the original study (N=97), (*ii*) using our annotation procedure (N=136), and (*iii*) the intersection between these gene sets (N=79). When doing so, we excluded regions mapping to the *HLA*-gene cluster and *IGH*-gene clusters (3 SNPs in total) as the original study did not report gene names but rather *HLA*-cluster and *IGH*-cluster. Nevertheless, our annotation procedure correctly assigned *HLA*-related genes and *IGH*-related genes with these SNPs. The number of intersecting genes was significantly higher than what could be expected by chance (p=0.03, based on one-tail p-value of binomial test, Table S2). For 6 SNPs, the gene annotated by our procedure did not match the gene assigned in the original study. Specifically, for 4/6 of these SNPs, we found significant eQTLs in blood (*rs60755019* with *ADCY10P1*, *rs7384878* with *PILRB, STAG3L5P, PMS2P1, GIGYF1*, and *EPHB4* genes, *rs56407236* with *FAM157C* gene, and *rs2526377* with *TRIM37* gene), while the original study reported the closest genes as most likely gene (*rs60755019* with *TREML2* gene,*rs7384878* with *SPDYE3* gene,*rs56407236* with *PRDM7* gene and *rs2526377* with *TSOAP1* gene). In addition, we annotated SNPs *rs76928645* and *rs139643391* to *SEC61G* and *WDR12* genes (closest genes), while the original study, using eQTL and TWAS in AD-relevant brain regions, annotated these SNPs to *EGFR* and *ICA1L/CARF* genes. While the latter two SNPs were likely mis-annotated in our procedure (due to specific datasets used for the annotation), our annotation of the former 4 SNPs seemed robust, and further studies will have to clarify the annotation of these SNPs. With the resulting list of input SNPs and (likely) associated genes, we probed the GWAS-Catalog and the datasets of structural variations for previously reported associations. We found a marked

enrichment in the GWAS-Catalog for Alzheimer's disease, family history of Alzheimer's disease, and lipoprotein measurement (**??**, Table S3 and Table S4). The results of this analysis are relevant to the user as they indicate other traits that were previously associated with the input SNPs. As such, they may suggest relationships between different traits, for example in our case study they suggest the involvement of cholesterol and lipid metabolism in AD, a known relationship.[43] Next, we searched for all structural variations in a region of $10kb$ surrounding the input SNPs, and we found that for 39/83 SNPs, a larger structural variations was present in the vicinity (Table S5), including the known VNTR (variable number of tandem repeats) in *ABCA7* gene,[44] and the known CNV (copy number variation) in *CR1, HLA-DRA*, and *PICALM* genes (Table S5).[45, 46, 47] This information may be particularly interesting for experimental researchers investigating the functional effect of SVs, and could be used to prioritize certain genomic regions. Because of the complex nature of large SVs, these regions have been largely unexplored, however technological improvements now make it possible to accurately measure SV alleles. We then performed our (sampling-based) gene-set enrichment analysis using Gene Ontology Biological Processes (GO:BP, default setting) and Reactome as gene-set sources, and Blood as tissue for the eQTL analysis. After averaging p-values across the number of iterations, we found N=132 significant pathways from Gene Ontology (FDR<1%) and N=4 significant pathways from Reactome (FDR<10%) (Figure 8.7 and Table S6). To facilitate the interpretation of the gene-set enrichment results, we clustered the significantly enriched terms from Gene Ontology based on a semantic similarity measure using *REVIGO* (Figure 8.3 B) and our term-based clustering approach (Figure 8.3C). Both methods are useful as they provide an overview of the most relevant biological processes associated with the input SNPs. Our clustering approach found five main clusters of GO terms (Figure 8.3C and Figure 8.8). We generated wordclouds to guide the interpretation of the set of GO terms of each cluster (Figure 8.3C). The five clusters were characterized by (1) trafficking and migration at the level of immune cells, (2) activation of immune response, (3) organization and metabolic processes, (4) beta-amyloid metabolism and (5) amyloid and neurofibrillary tangles formation and clearance (Figure 8.3C). All these processes are known to occur in the pathogenesis of Alzheime's disease from other previous studies.[42, 43, 48, 49] We observed that clusters generated by *REVIGO* are more conservative (*i.e.* only terms with a high similarity degree were merged) as compared to our term-based clustering which generates a higher-level overview. In the original study (Table S15 from [42]), the most significant gene sets related

to amyloid and tau metabolism, lipid metabolism and immunity. In order to calculate the extent of term overlap between results from the original study and our approach, we calculated semantic similarity between all pairs of significantly enriched terms in both studies. In addition to showing pairwise similarities between all terms, this analysis also shows how the enriched terms in the original study relate to the clusters found using our term-based approach. We observed patterns of high similarity between the significant terms in both studies (Figure 8.9). For example, terms in the "Activation of immune system" and the "Beta-amyloid metabolism" clusters (defined with our term-based approach), reported high similarities with specific subsets of terms from the original study. This was expected as these clusters represent the most established biological pathways associated with AD. The cluster "Trafficking of immune cells" had high similarity with a specific subset of terms from the original study, yet we also observed similarities with the *Activation of immune system* cluster, in agreement with the fact that these clusters were relatively close also in tree structure (Figure 8.3C). Similarly, high similarities were observed between the *Beta-amyloid metabolism* and the *Amyloid formation and clearance* clusters. Finally, the *Metabolic processes* had high degree of similarity with a specific subset of terms, but also with terms related to *Activation of immune system* cluster. Altogether, we showed that (*i*) enriched terms from the original study and our study had a high degree of similarity, and (*ii*) that the enriched terms of the original study resembled the structure of our clustering approach. The complete analysis of 83 genetic variants took about 30 minutes to complete.

8.4 Discussion

Despite the fact that many summary statistics of genetic studies have been publicly released, the integration of such a large amount of data is often difficult and requires specific tools and knowledge. Even simple tasks, such as the rapid interrogation of how well a certain genomic region associates with a specific trait or multiple traits can be frustrating and time consuming. Our main objective to develop *snpXplorer* was the need for an easy-to-use and user-friendly framework to explore, analyse and integrate outcomes of GWAS and other genetic studies. *snpXplorer* showed to be a robust tool that can support a complete GWAS analysis, from the exploration of specific regions of interest to the variant-to-gene annotation, gene-set enrichment analysis and interpretation of associated biological pathways. To our knowledge, the only existing web-server that offers a similar explorative framework

as *snpXplorer* is the GWAS-Atlas.[50] GWAS-Atlas was primarily developed as a database of publicly available GWAS summary statistics. It offers possibilities to visualise Manhattan and quantile-quantile (QQ) plots, to perform downstream analyses using MAGMA statistical framework, and to study genetic correlation between traits by means of LD score regression.[51, 52] However, *snpXplorer* was developed mainly for visualisation purposes, and thus incorporate multiple unique features such as the possibility to visualise multiple GWAS datasets simultaneously or to upload an external association dataset for additional comparisons with existing datasets. Moreover, *snpXplorer* annotates these visualisations with several genomic features such as structural variations, recombination rates, LD patterns and eQTLs. All the relevant information showed in *snpXplorer*, such as top SNP information, eQTL tables, LD tables and structural variants can be easily downloaded for further investigations. Further, we would like to stress the relevance of overlaying the GWAS results with structural variants found by third-generation sequencing. Such structural variations have already been shown to play a significant role for several traits, in particular for neurodegenerative diseases, and *snpXplorer* is thus far the only web-server where such information can be visualized in the context of GWAS summary statistics.[44, 45, 53, 54] We do acknowledge that for an in-depth functional annotation analysis of GWAS, the possibility of integrating additional ad-hoc information (such as eQTLs, sQTLs, eTWAS and sTWAS from specific disease-related regions) may improve the analysis, but such data is not always available, is time consuming and requires deep knowledge. Several online and offline tools have been developed with a similar goal, *e.g. SNPnexus, ANNOVAR, FUMA* and *Ensembl VEP*.[55, 56, 57, 58] Some of these tools are characterized by a larger list of annotation sources, for example implementing multiple tools for variant effect prediction (*e.g. SNPnexus, Ensembl VEP* or *ANNOVAR*), or more extensive pathway enrichment analyses at the tissue- and cell-type level (*e.g. FUMA*). We have shown that *snpXplorer* provides similar results in terms of annotation capabilities and gene-set enrichment analysis as compared to existing tools. Yet, snpXplorer has several unique features for the functional annotation section, such as the extensive interpretation analysis implemented in *REVIGO*, our term-based clustering approach and the word-cloud visualisation, or the possibility to associate multiple genes with each SNP during gene-set enrichment analysis. Moreover, *snpXplorer* development will continue by implementation of additional annotation sources and analyses. Altogether, we showed that *snpXplorer* is a promising functional annotation tool to support a typical GWAS analysis. As such, it has been

previously applied for the annotation and downstream analysis of genetic variants associated with Alzheimer's disease and human longevity.[49]

8.4.1 Future updates

For future updates, we plan to keep updated and increase the list of summary statistics available to be displayed in the exploration section. In its current version, the exploration section of *snpXplorer* requires the user to define a region of interest to look, while genome-wide comparisons are not considered. However, it is our intention to implement a genome-wide comparison across GWAS studies that, given a set of input GWASs and a significance threshold alpha, reports all SNPs with a $p<\alpha$ across the studies, allowing for a more rapid visualisation of overlapping SNP-associations. Moreover, we plan to increase the number of annotation sources and available options in the annotation section (for example, including methylation-QTL, protein-QTL and splicing-QTL). Finally, we are also working towards adding a framework to calculate weighted polygenic risk scores given a set of individuals' genotypes and a reference study to take variant effect-sizes from.

8.5 Availability

snpXplorer is an open-source web-server available at `https://snpxplorer.net`. Tutorial videos, full documentation and link to code are available in the *Help* page of the web-server. *snpXplorer* is running as from March-2020, was tested both within and outside our group, and runs steadily on both Unix and Windows most common browsers (Safari, Google Chrome, Microsoft Edge, Internet Explorer, and Firefox). For certain steps, *snpXplorer* does rely on external tools and sources (*e.g. REVIGO*), and consequently depends on their availability. Although discouraged, the tool can also be installed locally on your machine: additional information on how to do it are available in our github at `https://github.com/TesiNicco/SNPbrowser`, however, we note that for the stand-alone version additional files should be downloaded separately, for example, all summary statistics. *snpXplorer* requires R (v3.5+) and python (v3+) correctly installed and accessible in your system. *snpXplorer* uses the following R packages: *shiny, data.table, stringr, ggplot2, liftOver, colourpicker, rvest, plotrix, parallel, SNPlocs.Hsapiens.dfSNP144.GRCh37, lme4, ggsci, RColorBrewer, gprofiler2, GOSemSim, GO.db, org.Hs.eg.db, pheatmap, circlize, devtools, treemap, basicPlotteR, gwascat, GenomicRanges, rtracklayer, Homo.sapiens, BiocGenerics*, and the following python libraries: *re, werkzeug, robobrowser, pygosemsim, numpy, csv, networkx and sys.*

8.6 Acknowledgements and Funding

The authors declare no conflict of interests. This work was supported by Stichting Alzheimer Nederland (WE09.2014-03), Stichting Diorapthe, horstingstuit foundation, Memorabel (ZonMW projectnumber 733050814) and Stichting VUmc Fonds. **Conflict of interest:** all authors declare no conflict of interest.

8.7 Full author list and affiliations

Niccolo' Tesi,[1,2,3] Sven J. van der Lee,[1,2] Marc Hulsman,[1,2,3] ,[5] Henne Holstege [1,2,3] and Marcel J.T. Reinders [3]

[1] Alzheimer Centre, Department of Neurology, Amsterdam Neuroscience, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[2] Section Genomics of Neurodegenerative Diseases and Aging, Department of Clinical Genetics, Vrije Universiteit Amsterdam, Amsterdam UMC, Amsterdam, The Netherlands

[3] Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

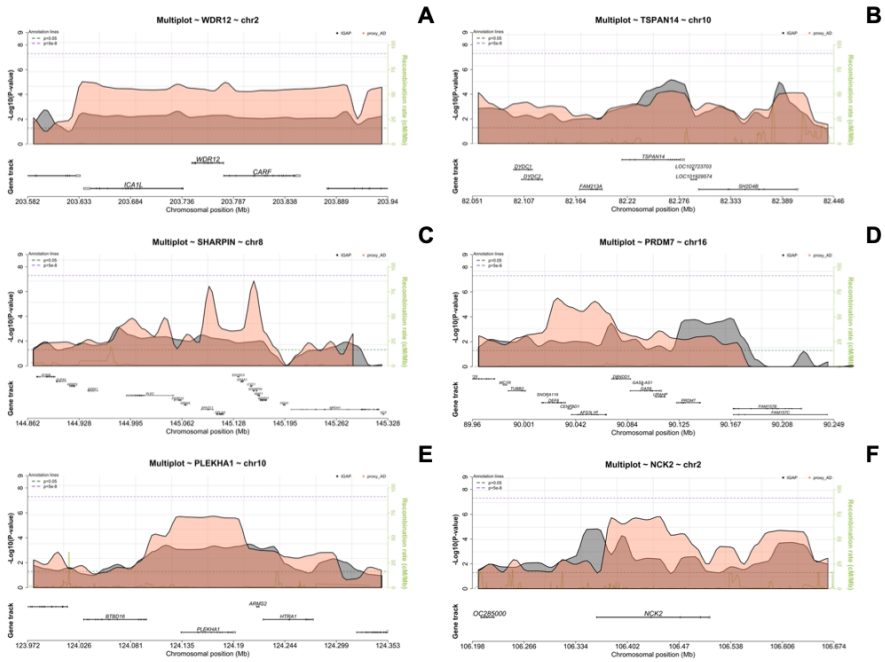


Figure 8.4: **Rapid exploration of novel SNP-associations in existing GWAS datasets. A.** The figure shows the region surrounding *WDR12* gene, for which a novel SNP-association was found in the case-study GWAS of Alzheimer's disease (AD). Two previous studies of AD are plotted, and show that suggestive association signals were present in earlier studies, yet the association did not reach genome-wide statistical significance, likely due to sample size. Similar plots show the regions surrounding *TSPAN14* (**B**), *SHARPIN* (**C**),*PRDM7* (**D**) ,*PLEKHA1* (**E**) and *NCK2* (**F**).

8.8 Supplementary Figures

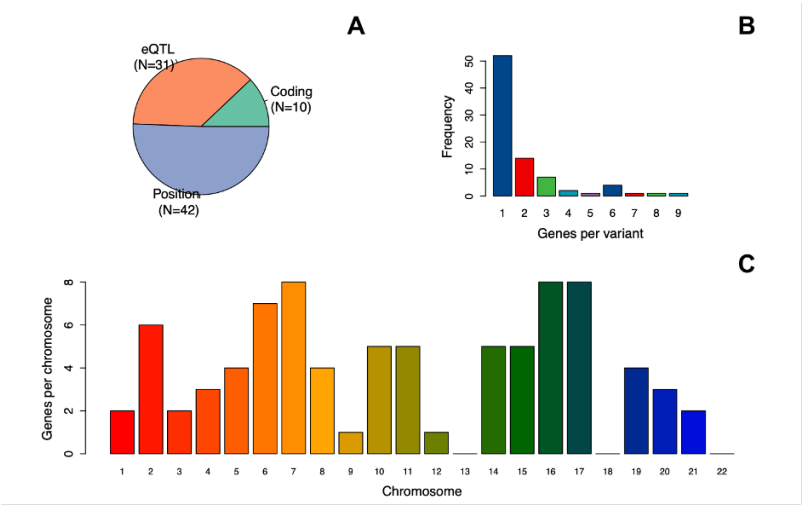


Figure 8.5: **Extreme AD cases vs. normal controls: E^{k}_{EA-NC}.Variant- gene mapping procedure. A.** The figure shows the type of genetic variants used as input, classified as coding, eQTL or annotated by their positions. **B.** The barplot shows the number of genes associated with each variant. **C.** The barplot shows the chromosomal distribution of all input variants.

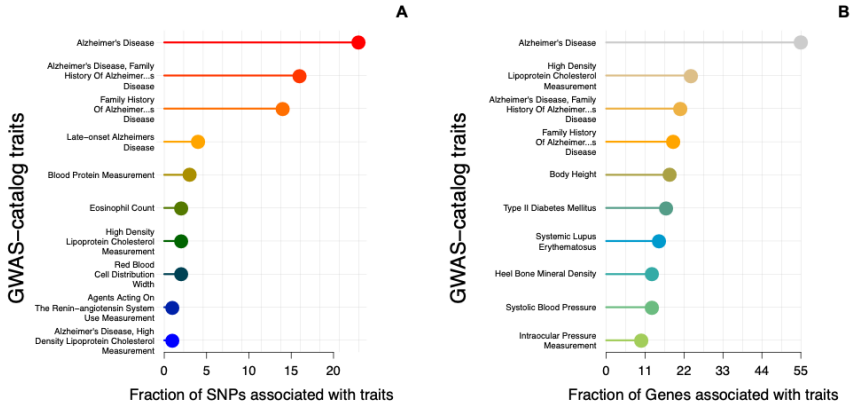


Figure 8.6: **Fraction of SNPs and Genes association with traits in the GWAS Catalog. A.** Number of input SNPs previously associated with traits in the GWAS catalog. **B.** Fraction of genes (associated with input SNPs) previously associated with traits in the GWAS Catalog.

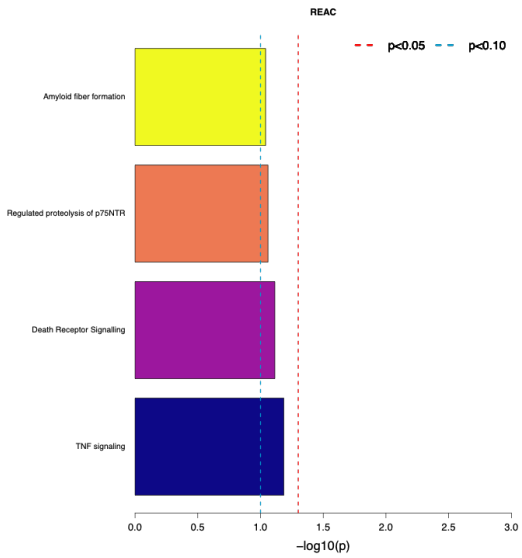


Figure 8.7: **Gene-set enrichment analysis.** The figure shows the barplot of the most significant pathways (FDR<10%) from Reactome.

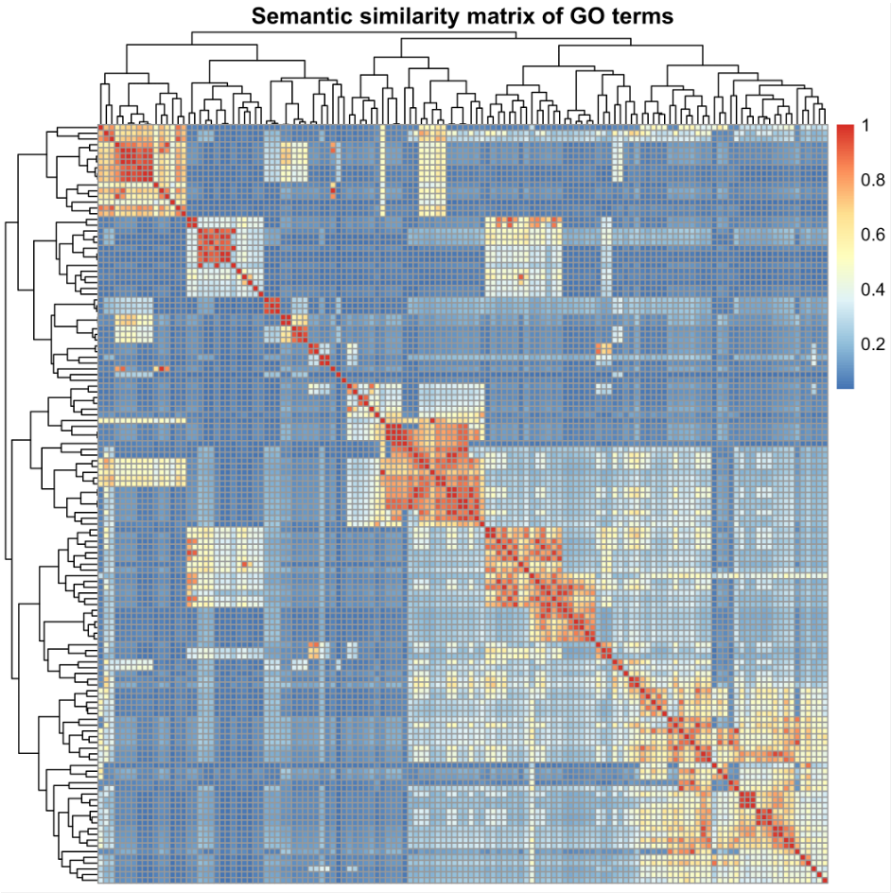


Figure 8.8: **Semantic similarity matrix.**The plot shows the semantic similarity matrix between all significantly enriched GO (Gene Ontology) biological processes terms (N=132). As semantic similarity, we used Lin.

Figure 8.9: **Comparison of gene-set enrichment results in the original study and using snpXplorer functional annotation section.**The heatmap shows the pairwise semantic similarity values between all significantly enriched terms in the original study (y-axis, N=92) and using *snpXplorer* functional annotation section (x-axis, N=132). The terms from our study (x-axis) are ordered based on their assigned cluster as a result of our term-based clustering approach. Large similarity patterns are visible in the heatmap, especially of terms (from the original study) mapping to *Activation of immune response* cluster (red cluster) and to *Beta-amyloid metabolism* cluster (green cluster). Some enriched terms mapping to *Trafficking of immune cells* (black cluster) had high similarity with *Activation of immune response* (purple cluster) cluster, and some terms mapping to *Amyloid formation and clearance* had high similarity with *Beta-amyloid metabolism*, resembling the structure of the tree constructed in our study. The remaining *Metabolic processes* cluster (blue cluster) had high similarity with a specific subset of enriched terms, but we also observed high similarity with the *Activation of immune system* cluster.

8.9 Supplementary Tables

Supplementary Tables can be accessed by scanning the following code or accessing the journal's website here.

References

[1] Annalisa Buniello et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". In: *Nucleic Acids Research* 47 (D1 Jan. 2019), pp. D1005–D1012. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gky1120.

[2] Alzheimer Disease Genetics Consortium (ADGC), et al. "Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing". In: *Nature Genetics* 51.3 (Mar. 2019), pp. 414–430. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-019-0358-2.

[3] Iris E. Jansen et al. "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk". In: *Nature Genetics* 51.3 (Mar. 2019), pp. 404–413. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-018-0311-9.

[4] Nana Matoba et al. "Common genetic risk variants identified in the SPARK cohort support DDHD2 as a candidate risk gene for autism". en. In: *Translational Psychiatry* 10.1 (Dec. 2020), p. 265. ISSN: 2158-3188. DOI: 10.1038/s41398-020-00953-9.

[5] MDD Working Group of the Psychiatric Genomics Consortium et al. "Minimal phenotyping yields genome-wide association signals of low specificity for major depression". en. In: *Nature Genetics* 52.4 (Apr. 2020), pp. 437–447. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-020-0594-5.

[6] Dina Vojinovic et al. "Genome-wide association study of 23,500 individuals identifies 7 loci associated with brain ventricular volume". en. In: *Nature Communications* 9.1 (Dec. 2018), p. 3945. ISSN: 2041-1723. DOI: 10.1038/s41467-018-06234-w.

[7] Tove Fall et al. "Genome-wide association study of coronary artery disease among individuals with diabetes: the UK-Biobank". In: *Diabetologia* 61.10 (Oct. 2018), pp. 2174–2179. ISSN: 0012-186X, 1432-0428. DOI: 10.1007/s00125-018-4686-z.

[8] the Million Veteran Program et al. "Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits". en. In: *Nature Genetics* 50.10 (Oct. 2018), pp. 1412–1425. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-018-0205-x.

[9] Loic Yengo et al. "Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry". en. In: *Human Molecular Genetics* 27.20 (Oct. 2018), pp. 3641–3649. ISSN: 0964-6906, 1460-2083. DOI: 10.1093/hmg/ddy271.

[10] Vincenzo Forgetta et al. "Rare Genetic Variants of Large Effect Influence Risk of Type 1 Diabetes". en. In: *Diabetes* 69.4 (Apr. 2020), pp. 784–795. ISSN: 0012-1797, 1939-327X. DOI: 10.2337/db19-0831.

[11] The GenOMICC Investigators et al. "Genetic mechanisms of critical illness in Covid-19". en. In: *Nature* (Dec. 2020). ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-03065-y.

[12] Yong-Fei Wang et al. "Identification of 38 novel loci for systemic lupus erythematosus and genetic heterogeneity between ancestral groups". en. In: *Nature Communications* 12.1 (Dec. 2021), p. 772. ISSN: 2041-1723. DOI: 10.1038/s41467-021-21049-y.

[13] FinnGen et al. "An expanded analysis framework for multivariate GWAS connects inflammatory biomarkers to functional variants and disease". en. In: *European Journal of Human Genetics* 29.2 (Feb. 2021), pp. 309–324. ISSN: 1018-4813, 1476-5438. DOI: 10.1038/s41431-020-00730-8.

[14] Yi Han et al. "Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma". en. In: *Nature Communications* 11.1 (Dec. 2020), p. 1776. ISSN: 2041-1723. DOI: 10.1038/s41467-020-15649-3.

[15] kConFab Investigators et al. "Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses". en. In: *Nature Genetics* 52.6 (June 2020), pp. 572–581. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-020-0609-2.

[16] FinnGen et al. "Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells". en. In: *Nature* 586.7831 (Oct. 2020), pp. 769–775. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-2786-7.

[17] Peter N. Fiorica et al. "Multi-ethnic transcriptome-wide association study of prostate cancer". en. In: *PLOS ONE* 15.9 (Sept. 2020). Ed. by Amanda Ewart Toland, e0236209. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0236209.

[18] Sara R. Rashkin et al. "Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts". en. In: *Nature Communications* 11.1 (Dec. 2020), p. 4423. ISSN: 2041-1723. DOI: 10.1038/s41467-020-18246-6.

[19] Paul RHJ Timmers et al. "Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances". In: *eLife* 8 (Jan. 2019). ISSN: 2050-084X. DOI: 10.7554/eLife.39856.

[20] Perline A. Demange et al. "Investigating the genetic architecture of noncognitive skills using GWAS-by-subtraction". en. In: *Nature Genetics* 53.1 (Jan. 2021), pp. 35–44. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-020-00754-2.

[21] Regeneron Genetics Center et al. "MEPE loss-of-function variant associates with decreased bone mineral density and increased fracture risk". en. In: *Nature Communications* 11.1 (Dec. 2020), p. 4093. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17315-0.

[22] Despoina Manousaki et al. "Genome-wide Association Study for Vitamin D Levels Reveals 69 Independent Loci". en. In: *The American Journal of Human Genetics* 106.3 (Mar. 2020), pp. 327–337. ISSN: 00029297. DOI: 10.1016/j.ajhg.2020.01.017.

[23] W. J. Kent et al. "The Human Genome Browser at UCSC". en. In: *Genome Research* 12.6 (May 2002), pp. 996–1006. ISSN: 1088-9051. DOI: 10.1101/gr.229102.

[24] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *American Journal of Human Genetics* 81.3 (Sept. 2007), pp. 559–575. ISSN: 0002-9297. DOI: 10.1086/519795.

[25] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (Oct. 2015), pp. 68–74. ISSN: 1476-4687. DOI: 10.1038/nature15393.

[26] Nuala A. O'Leary et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic Acids Research* 44 (D1 Jan. 2016), pp. D733–745. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1189.

[27] The International HapMap Consortium. "A second generation human haplotype map of over 3.1 million SNPs". In: *Nature* 449.7164 (Oct. 2007), pp. 851–861. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature06258.

[28] Mark J. P. Chaisson et al. "Multi-platform discovery of haplotype-resolved structural variation in human genomes". en. In: *Nature Communications* 10.1 (Dec. 2019), p. 1784. ISSN: 2041-1723. DOI: 10.1038/s41467-018-08148-z.

[29] Peter A. Audano et al. "Characterizing the Major Structural Variant Alleles of the Human Genome". en. In: *Cell* 176.3 (Jan. 2019), 663–675.e19. ISSN: 00928674. DOI: 10.1016/j.cell.2018.12.019.

[30] Jasper Linthorst et al. "Extreme enrichment of VNTR-associated polymorphicity in human subtelomeres: genes with most VNTRs are predominantly expressed in the brain". en. In: *Translational Psychiatry* 10.1 (Dec. 2020), p. 369. ISSN: 2158-3188. DOI: 10.1038/s41398-020-01060-5.

[31] GTEx Consortium. "The Genotype-Tissue Expression (GTEx) project". In: *Nature Genetics* 45.6 (June 2013), pp. 580–585. ISSN: 1546-1718. DOI: 10.1038/ng.2653.

[32] Gil Stelzer et al. "The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses". en. In: *Current Protocols in Bioinformatics* 54.1 (June 2016). ISSN: 1934-3396, 1934-340X. DOI: 10.1002/cpbi.5.

[33] Jie Zheng et al. "LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis". eng. In: *Bioinformatics (Oxford, England)* 33.2 (2017), pp. 272–279. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btw613.

[34] Philipp Rentzsch et al. "CADD: predicting the deleteriousness of variants throughout the human genome". In: *Nucleic Acids Research* 47 (D1 Jan. 2019), pp. D886–D894. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gky1016.

[35] Uku Raudvere et al. "g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)". In: *Nucleic Acids Research* 47 (W1 July 2019), W191–W198. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz369.

[36] M. Ashburner et al. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium". In: *Nature Genetics* 25.1 (May 2000), pp. 25–29. ISSN: 1061-4036. DOI: 10.1038/75556.

[37] Minoru Kanehisa et al. "KEGG: integrating viruses and cellular organisms". en. In: *Nucleic Acids Research* 49.D1 (Jan. 2021), pp. D545–D551. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkaa970.

[38] Bijay Jassal et al. "The reactome pathway knowledgebase". en. In: *Nucleic Acids Research* (Nov. 2019), gkz1031. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkz1031.

[39] Alexander R. Pico et al. "WikiPathways: pathway editing for the people". eng. In: *PLoS biology* 6.7 (July 2008), e184. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0060184.

[40] Fran Supek et al. "REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms". In: *PLoS ONE* 6.7 (July 2011). Ed. by Cynthia Gibas, e21800. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021800.

[41] Bridget T. McInnes and Ted Pedersen. "Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text". In: *Journal of Biomedical Informatics* 46.6 (Dec. 2013), pp. 1116–1124. ISSN: 15320464. DOI: 10.1016/j.jbi.2013.08.008.

[42] Céline Bellenguez et al. *New insights on the genetic etiology of Alzheimer's and related dementia*. en. preprint. Neurology, Oct. 2020. DOI: 10.1101/2020.10.01.20200659.

[43] J. Hardy et al. "Pathways to Alzheimer's disease". In: *Journal of Internal Medicine* 275.3 (Mar. 2014), pp. 296–303. ISSN: 09546820. DOI: 10.1111/joim.12192.

[44] Arne De Roeck, Christine Van Broeckhoven, and Kristel Sleegers. "The role of ABCA7 in Alzheimer's disease: evidence from genomics, transcriptomics and methylomics". In: *Acta Neuropathologica* 138.2 (Aug. 2019), pp. 201–220. ISSN: 0001-6322, 1432-0533. DOI: 10.1007/s00401-019-01994-1.

[45] N Brouwers et al. "Alzheimer risk associated with a copy number variation in the complement receptor 1 increasing C3b/C4b binding sites". en. In: *Molecular Psychiatry* 17.2 (Feb. 2012), pp. 223–233. ISSN: 1359-4184, 1476-5578. DOI: 10.1038/mp.2011.24.

[46] Shanker Swaminathan et al. "Analysis of copy number variation in Alzheimer's disease in a cohort of clinically characterized and neuropathologically verified individuals". eng. In: *PloS One* 7.12 (2012), e50640. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0050640.

[47] Kinga Szigeti et al. "Genome-wide scan for copy number variation association with age at onset of Alzheimer's disease". eng. In: *Journal of Alzheimer's disease: JAD* 33.2 (2013), pp. 517–523. ISSN: 1875-8908. DOI: 10.3233/JAD-2012-121285.

[48] Caroline Van Cauwenberghe, Christine Van Broeckhoven, and Kristel Sleegers. "The genetic landscape of Alzheimer disease: clinical implications and perspectives". In: *Genetics in Medicine* 18.5 (May 2016), pp. 421–430. ISSN: 1098-3600, 1530-0366. DOI: 10.1038/gim.2015.117.

[49] Niccolò Tesi et al. "Immune response and endocytosis pathways are associated with the resilience against Alzheimer's disease". In: *Translational Psychiatry* 10.1 (Dec. 2020), p. 332. ISSN: 2158-3188. DOI: 10.1038/s41398-020-01018-7.

[50] Kyoko Watanabe et al. "A global overview of pleiotropy and genetic architecture in complex traits". eng. In: *Nature Genetics* 51.9 (Sept. 2019), pp. 1339–1348. ISSN: 1546-1718. DOI: 10.1038/s41588-019-0481-0.

[51] Christiaan A. de Leeuw et al. "MAGMA: Generalized Gene-Set Analysis of GWAS Data". In: *PLOS Computational Biology* 11.4 (Apr. 2015). Ed. by Hua Tang, e1004219. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004219.

[52] Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature Genetics* 47.3 (Mar. 2015), pp. 291–295. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.3211.

[53] Rubika Balendra and Adrian M. Isaacs. “C9orf72-mediated ALS and FTD: multiple pathways to disease”. en. In: *Nature Reviews Neurology* 14.9 (Sept. 2018), pp. 544–558. ISSN: 1759-4758, 1759-4766. DOI: 10.1038/s41582-018-0047-2.

[54] Douglas R. Langbehn et al. “CAG-repeat length and the age of onset in Huntington disease (HD): A review and validation study of statistical approaches”. en. In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 153B.2 (Mar. 2010), pp. 397–408. ISSN: 15524841. DOI: 10.1002/ajmg.b.30992.

[55] Jorge Oscanoa et al. “SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update)”. en. In: *Nucleic Acids Research* 48.W1 (July 2020), W185–W192. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkaa420.

[56] K. Wang, M. Li, and H. Hakonarson. “ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data”. en. In: *Nucleic Acids Research* 38.16 (Sept. 2010), e164–e164. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkq603.

[57] Kyoko Watanabe et al. “Functional mapping and annotation of genetic associations with FUMA”. In: *Nature Communications* 8.1 (Dec. 2017), p. 1826. ISSN: 2041-1723. DOI: 10.1038/s41467-017-01261-5.

[58] William McLaren et al. “The Ensembl Variant Effect Predictor”. en. In: *Genome Biology* 17.1 (Dec. 2016), p. 122. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0974-4.

9. General discussion

Key findings

The overall objective of this thesis was to investigate the genetic factors underlying maintained cognitive health until extreme ages, in the context of Alzheimer's disease (AD) and extreme human longevity. Here, we provide a summary of our key findings:

- Cognitively healthy centenarians have a lower frequency of common variants that increase the risk to develop AD and have a higher frequency of variants that protect against the disease.
- The risk to develop, or being resilient against, AD changes on a pathway basis: we found that the polygenic risk of variants involved in immune response and endocytosis processes associated the most with the resilience against AD
- The majority of alleles that increase the risk of AD, negatively affect lifespan. While for most alleles the negative effect on lifespan is explained through the increased risk of AD, a subset of alleles appeared to be beneficial for healthy aging, and as such, decreased the risk of AD
- A polygenic risk score of longevity comprising 330 variants associated with becoming a cognitively healthy centenarian, led up to 4-year difference in survival and functionally associated with hallmarks of longevity
- We combined previous GWAS of clinical AD and AD by-proxy in the largest genome-wide meta-analysis of AD to date, which led to the discovery of six novel genetic variants that influence the risk of AD, expanding the knowledge of the genetic landscape of AD
- In the largest genome-wide association meta-analysis of longevity to date, we discovered one variant in addition to *APOE* variants near *GPR78* gene to be significantly associated with longevity, and identified a shared genetic architecture of health and longevity
- We have developed *snpXplorer*, an open-source web application that allows the exploration of GWAS association results, as well as the functional annotation and the pathway enrichment analysis of any given set of genetic variants

9.1 General discussion

The study of individuals with extreme longevity has enabled the discovery of environmental and genetic factors that impact lifespan.[1, 2] However, with an increasing fraction of individuals reaching old age, the next challenge is to determine which factors are associated with a *healthy* lifespan. Reaching 100 years is only satisfying when chronic illnesses and cognitive decline due to dementia can be escaped as much as possible. The 100-plus Study and other centenarian studies around the world have demonstrated that although rare, this is possible. About ~10% of all centenarians can be considered to be in good cognitive and physical conditions.[3, 4] This subgroup of individuals may have specific characteristics that protect or delay the onset of cognitive impairment and other age-related diseases, which emphasizes the need to explore the underlying mechanisms that maintained their cognitive health. In this chapter, we interpret the most important findings presented in this thesis, including the exploration of genetic factors in cognitively healthy centenarians. We will further discuss challenges, limitations, and future perspectives.

9.2 Genetic factors influencing resilience against Alzheimer's disease

Resilience is the process that allows individuals to withstand adverse conditions.[5] In the context of Alzheimer's disease (AD) research, this refers to the ability to escape the development of clinical symptoms of AD, while being exposed to risk factors that would normally result in cognitive decline.[6] Given the high heritability of AD, such resilience mechanisms may also be genetically encoded. However, while cognitive resilience is an active research field in psychology and psychiatry, the extent of the role of the genetically encoded resilience, in AD and other age-related disease, is largely unexplored. In this thesis, we investigated for the first time the role of genetic factors underlying the resilience against Alzheimer's disease in individuals that reached extremely old ages without suffering from dementia. In chapter 2 and chapter 3, we studied the frequency of AD-associated genetic variants in cognitively healthy agers, middle-aged population controls, and relatively young AD cases.[7, 8] These individuals together cover the entire cognitive spectrum, with cognitively healthy agers and AD cases representing the two extremes. Therefore, using the healthy population subjects as the middle point, we were able to study, on the one end, the risk to develop AD (in a comparison of young AD cases and population controls),

and, on the other end, the resilience against AD (by comparing population controls and cognitively healthy agers). Comparing the two extremes in the cognitive spectrum, we observed a remarkable average 2-fold enrichment in the variant effect-sizes compared to expected effect-sizes from published GWAS. At the level of the single variant, the effect-size increase was as high as 6-fold. Importantly, this enrichment was mainly driven by the cognitively healthy centenarians. This means that cognitively healthy agers are depleted with genetic variants associated with increased AD-risk compared to the general population, and that the study of individuals with extreme phenotypes is protifable for the research of genetic factors associated with AD. However, our research raises additional questions regarding the nature of extreme cases, *e.g.* the existence of rare, undiscovered genetic variations that may explain the early onset of the disease. Interestingly, the degree of depletion/enrichment of the different AD variants in the centenarians was not constant but appeared to cluster on a pathway basis. In chapter 3, we focused on the major pathways thought to underly AD pathogenesis, and we combined the effect of multiple variants in Polygenic Risk Scores (PRSs) and pathway-specific PRSs. We showed that a PRS including all AD-associated variants significantly associated with both the increased risk and the resilience against AD. Furthermore, we found that the escaping AD was genetically encoded by variants associated with immune-related processes, even after excluding the large effect of *APOE*-associated variants. Although not specifically on AD, previous animal and human studies have highlighted the relationship between immune response and psychological resilience in humans.[5] Genetic variants involved in immune-related mechanisms may also contribute to such effect, eventually making resilient individuals able to better recover from inflammation-induced stressors.[5] Altogether, we observed that achieving extreme ages with maintained cognitive health is encoded, at least in part, by genetic factors that are specifically involved in immune-related mechanisms. This suggests that modulating immune-related pathways may be a feasible strategy to prevent AD.

9.3 APOE alleles in cognitively healthy centenarians

The strongest genetic factors associated with increased AD-risk, protection against AD, and human longevity, are the $\epsilon 2$ and $\epsilon 4$ alleles in the *APOE* gene.[9, 10, 11, 12] Despite >20 years of research, the mechanisms by which *APOE* affects molecular pathways of AD is not completely understood. Although more research will eventually lead to a deeper understanding

of *APOE* functions, it seems that the *APOE* gene is involved in all major known AD-associated pathways (β-amyloid metabolism, lipid and cholesterol metabolism, and immunity), which we also showed in chapter 3 and chapter 4. Likely due to the combination of effect on both AD and longevity, our cohort of cognitively healthy centenarians is remarkably enriched for the $\epsilon 2$ allele and depleted for the $\epsilon 4$ allele. That is, while the frequency of $\epsilon 2$ and ϵ alleles is respectively 8% and 16% in the general population, in AD is about 4% and 43%, and in our centenarians the frequencies were 16% and 8%.[7] For this reason, it was necessary to correct the analyses in chapter 3 and chapter 5 according to carriership of *APOE* variants: after correction, we still observed a significant difference in the PRS and pathway-PRS between centenarians and population subjects, on both AD and longevity. Of note: in our cohort of centenarians, *N*=47 carried at least one *APOE* $\epsilon 4$ allele, and a single centenarian was homozygous for *APOE* $\epsilon 4$ ($\epsilon 4/\epsilon 4$); compared to those who did not carry a deleterious *APOE* $\epsilon 4$ allele, the carriers reported a significantly lower polygenic risk score for AD (excluding *APOE* variants), suggesting that even the negative effects of *APOE* $\epsilon 4/\epsilon 4$ genotype can be balanced out through other (protective) variants.

9.4 The aging effect of AD-associated variants

In chapter 2 and chapter 3, we showed that cognitively healthy centenarians are genetically protected against AD, largely due to genetic variants involved in immune-related processes. As cognitively healthy centenarians are both free from dementia and extremely old, this suggests that the etiology of AD and healthy lifespan might overlap across these biological pathways. In chapter 4 and chapter 5, we have studied in depth to what extent genetic variants associated with AD and other age-related diseases are related to extreme human longevity. The current view of the genetics of longevity is that it depends on a depletion of genetic elements associated with an increased risk of age-related diseases.[1, 13, 14] Therefore, given the large prevalence of AD at old ages and the increased mortality due to the disease, one would expect that variants increasing the risk of AD, would negatively affect lifespan. In accordance with this hypothesis, we showed that the majority of alleles increasing the risk of AD, negatively affected longevity. However, we identified different trajectories of effect on healthy aging of AD-associated genetic variants. Firstly, genetic variants that increase the risk of AD and as a consequence, they harm longevity; these genetic variants are likely the variants with the *"purest"* AD-effect (*e.g. CR1, CD33, BIN1, PICALM*,

and *MS4A6A*).[15] Secondly, genetic variants that primarily affect longevity, suggesting that the negative effect on lifespan contributes to the increased risk to develop AD, but potentially also other diseases. For example, in this group of variants we find the non-synonymous variant in *PLCG2* which was recently found to be protective against other conditions than AD, but also genetic variants in *APOE, SHARPIN, IQCK, PRKD3, CD2AP, HLA,* and *SPI1* genes, which were previously associated with respiratory system disease, cardiovascular diseases, autoimmune disorders, and cancer.[16, 17, 18, 19, 20, 21, 22] In line with the findings from chapter 2 and chapter 3, this group of genetic variants was also strongly enriched at the functional level for immune system response and endocytosis, which align to known hallmarks of aging. Before us, only one study investigated the relationship between AD-associated variants and longevity: they did not have access to extremely old (and healthy) individuals, used a relatively small sample size and did test only a small number of variants.[23] Apart from *APOE*, they could not find any significant effect on the longevity of AD-associated variants.[23] Likely, due to a larger number of variants that we studied as well as the extreme phenotype of the cognitively healthy centenarians, we were the first to investigate genetic factors and pathways associated with the resilience against AD, and to show potential pleiotropic effects on longevity of genetic variants associated with AD.

9.5 Genetic predisposition to extreme longevity

Whilst the analysis in chapter 4 was limited to AD, in chapter 5 we used the summary statistics from the largest GWAS on parental longevity to date to prioritize the genetic factors that have the largest effect on becoming a cognitively healthy centenarian.[24] This resulted in a PRS comprising 330 genetic variants that not only associated with becoming a cognitively healthy centenarian but interestingly, associated with up to a 4-year difference in survival in an independent cohort of healthy, middle-aged individuals. A previous study in literature showed that a PRS of parental longevity was significantly associated with survival, and we validate these results in a cohort of cognitively healthy centenarians. In line with expectations, the majority of the genetic variants included in the PRS were previously associated with several age-related conditions, including cardiovascular and autoimmune diseases, as well as cancer. Interestingly, we found suggestive evidence of the compensation-effect between the PRS and the *APOE* $\epsilon 4$ allele, as individuals who carried an *APOE* $\epsilon 4$ allele and had a high longevity-PRS

survived longer than those that did not carry an *APOE* ϵ4 allele and had a lower PRS. Such compensation mechanisms for the *APOE* ϵ4 allele were previously observed in a number of studies of AD, and although the results did not strongly replicate across different studies, several variants (*i.e. rs5882* in the *CETP* gene, *rs10553596* in the *CASP7* gene, and *rs4934* in the *SERPINA3* gene) were reported to exhibit buffering effects with respect to the presence of the *APOE* ϵ4 allele.[25, 26] In functional terms, aging is associated with a high degree of inflammation, cellular stress, as well as reduced capacity of cell-differentiation and development.[1, 13] We showed that the variants included in the PRS were functionally enriched for oxidative stress reduction and cell/tissue differentiation. Altogether, we confirmed in a cohort of cognitively healthy agers, that the human lifespan is influenced by a constellation of genetic variations distributed along the genome, likely acting to diminish the risk of age-related diseases and to balance out the alterations that physiologically take place in the aging individual. Further studies concerning the downstream effects of such genetic variants and their implications at the gene- and pathway-level may be of interest to the development of anti-aging drugs.

9.6 Extreme phenotypes in GWAS

In the previous chapters, we have shown that using extreme phenotypes represent an added value for the genetic research of complex polygenic traits. While we focused mainly on common and low-frequency genetic variants (minor allele frequency >1%), before this thesis, the use of extreme phenotypes in genetic studies was mainly applied to discover rare, causative, mutations responsible for various age-related diseases.[27, 16, 14, 17, 18] In such a setting, usually, extreme cases (for example, individuals that manifested clinical symptoms of a disease at a younger age or with extreme clinical manifestations) were compared to healthy controls. Only a handful of studies included both extremes of a disease spectrum, and showed, similarly to us, a relative increase in effect-size. For example, centenarians were labeled *"super-controls"* for the study of genetic factors underlying diabetes and possibly other age-related diseases.[28] In chapter 2 and chapter 3, we, for the first time, reported similar findings in the context of AD. The inclusion of individuals with extreme phenotypes in genetic studies led to an increase in the variant effect-sizes, which translates to higher statistical power to detect significant associations. While the main drawback of extreme phenotypes is their availability, which precludes the possibility to gather very

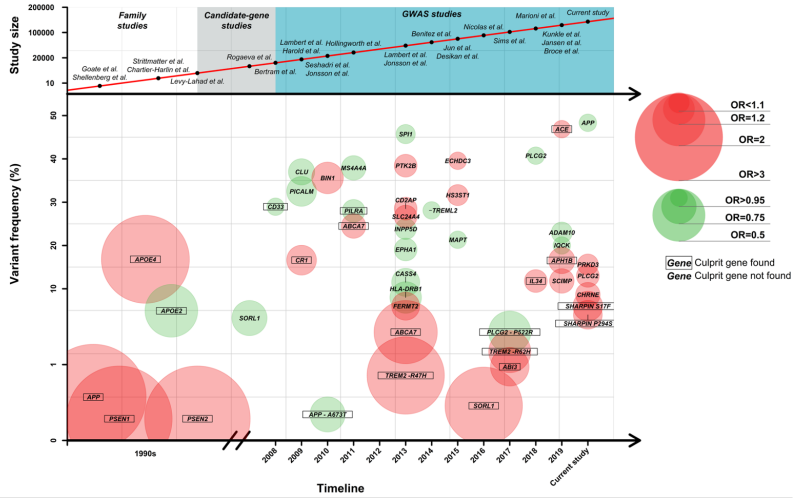


Figure 9.1: **The updated genetic landscape of Alzheimer's disease from genetic studies.** The top panel of the figure shows the course of the genetic research of AD, from early 1990's with the discovery of the causal mutations in *APP, PSEN1* and *PSEN2*, until nowadays, with the discovery of ~40 common genetic variants significantly associated with AD. However, the research of new genetic determinants is continuing, and larger GWAS will likely find association of additional rare and common genetic variants with AD.

large sample sizes, our findings might have methodological implications in terms of study design and sample selection for future genetic studies of AD, other age-related diseases, and longevity.[29]

9.7 Large collaborative efforts make the difference

With the knowledge of the protective elements retained in the genome of cognitively healthy agers, we sought to join other researchers in chapter 6 and chapter 7 by participating in two of the largest, collaborative genome-wide meta-analyses of AD and longevity to date. These collaborative efforts were important to further improve the knowledge of the genetic landscape underlying these traits. For AD, our study resulted in the discovery of 6 additional genetic variants associated with an increased risk of AD, which raised the total number of genetic variants associated with AD to 41.[30] This number will likely increase in the future due to an increased number of samples compared. [31] Although the number of centenarians included was

small compared to the total number of controls, we note that the majority of all the newly discovered variants associated significantly and in the correct direction in a comparison of centenarians and AD cases within our cohort only (Figure 9.1). Another practical implication of our meta-analysis of AD concerned the effectiveness of the PRS in a clinical setting. To date, despite the strong predictive effect exerted by PRS, it is currently not used as an additional tool for AD diagnosis.[30, 32, 33] Nevertheless, the identification, before onset of AD, of individuals at the highest risk of developing AD or individuals who are compromised in a certain pathway, is a promising tool in a diagnostic setting, and for the development and the application of preventative personalized treatment strategies.[34, 35] Unlike AD, genetic factors that influence longevity are more complex to study. For example, the phenotype definition is not clear, and which individuals should be used as cases or controls is not easily definable. In chapter 7, we and our collaborators defined a novel, significantly improved classification method of cases and controls based on the age at $90/99^{th}$ and 60^{th} percentile of survival probability per country. This resulted in a harmonization of the criteria to define cases and controls across different cohorts and countries, facilitating collaborative efforts. In terms of findings, we confirmed the longevity-association of variants in the *APOE* gene and propose the association of a novel variant (*rs7676745*) near the *GPR78* gene.[14] This variant was previously associated with psychiatric disorders, and its association was significant and in the correct direction in our study alone. Besides, through genetic correlation analyses, we showed that the genetic architecture between health and longevity overlaps, which finds accordance with previous studies, and our results from chapter 5. Nevertheless, the major challenge to study human longevity is the lack of a large sample size, as such individuals are rare and need to be approached individually.[4] One way out is to combine case-controls studies with by-proxy studies, similarly to what we have done in chapter 6 for AD, which then may lead to a better understanding of the genetic landscape of extreme human longevity.

9.8 Towards an updated disease model of AD

Our results in chapter 6 had important implications in terms of AD biology. On the one end, we identified a common genetic variation in the *APP* (Amyloid Precursor Protein) gene that increased the risk of AD, which enforces the evidence that *APP* processing is an important risk factor of AD, next to other strong risk factors, and, moreover, links the sporadic form of AD

with the autosomal dominant form. On the other end, we added on the growing importance of the immune system in AD development, as multiple new genetic loci pointed to genes involved in immune-signaling (*e.g. PLCG2, SHARPIN, CD33,* and *IL34*). This is in line with what we showed in chapter 3 where we have quantified the contribution of each AD-associated pathway to the total polygenic risk of AD, showing that, excluding *APOE* variants, ∼65% of the total risk of AD is due to variants involved in immune response and endocytosis pathways (Figure 9.2). These latter two biological pathways have been, in fact, proposed as potential central pathways whose dysfunction may trigger the molecular cascade leading to β-amyloid and tau accumulation, and synaptic and neuronal damage. Physiologically, the resident immune cells within the brain are involved, among other functions, in the clearing of cellular debris including aggregated β-amyloid peptides. However, the brain immune response may not react sufficiently or vice versa, it may react too strongly, possibly starting a molecular cascade typical of AD that leads to cognitive decline.[36] Genetic variants involved in immune response may modify how cellular debris and β-amyloid deposits in the brain are recognized, captured, and cleared (Figure 9.2).[37] To this end, the post-mortem analysis of the brains of cognitively healthy centenarians showed that, despite being cognitively healthy, these centenarians are not free from the typical neuropathological hallmarks of AD (accumulation of β-amyloid plaques and neurofibrillary tangles), yet, these are not severe enough to cause cognitive decline.[38] This capacity of monitoring the neuropathological hallmarks of the disease may be maintained in cognitively healthy centenarians due to a lower vulnerability of age-related decline, possibly as a result of a depletion of deleterious genetic variants and a concurrent enrichment of protective genetic elements, as we saw in chapter 2, chapter 3, and chapter 5. An alternative theory explaining AD development identifies in the endosomal trafficking pathway the trigger that starts the pathological events associated with Alzheimer's disease, culminating in β-amyloid and tau pathology, neuronal and microglial dysfunction. The endosomal trafficking pathway is a crucial molecular pathway that regulates the trafficking of intracellular and extracellular proteins and lipids (including the toxic β-amyloid peptides), and allows to direct them to other cellular compartments, to sort them for degradation, or to recycle them to the extracellular space. [39] An abnormal trafficking of endosomes, together with a potential dysfunction within the clearence system, may be the first step leading to amyloid, tau deposition, and synaptic and neuronal loss (Figure 9.2).[40]

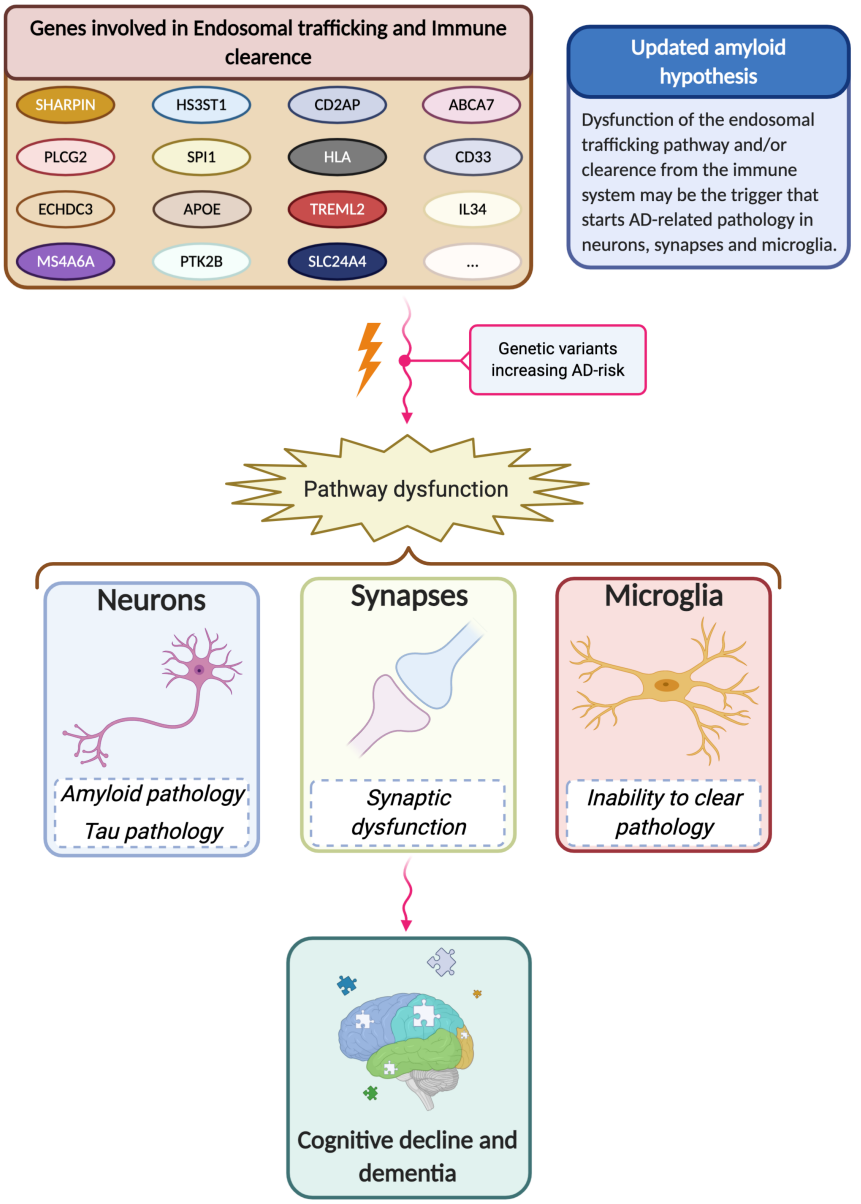


Figure 9.2: **An updated hypothetical model of Alzheimer's disease.**

9.9 Interpretation of GWAS

One of the major complexities in GWAS studies is the functional interpretation of the effect that a certain genetic variant has at the gene-, protein- and pathway-level, because the majority of genetic variants analyzed through GWAS relies in non-coding regions. In chapter 3, we approached this complexity by using published variant-gene associations as found in two previous GWASs of AD, which implemented different techniques to fine-map the variant-gene association.[41, 42] However, this approach was limited to AD. For other traits, several tools have been developed to address the variant-gene association,[43, 44] including the approach that we propose in chapter 4, that we further improved in chapter 5, and that evolved in the web-server application that we presented in chapter 8. Our approach was to use multiple resources, such as tissue-specific RNA expression and predicted variant effect, to allow multiple genes to associate with each variant, depending on the annotation certainties.[45, 46] However, because neighboring genes are often functionally related, allowing multiple genes to associate with a variant could result in an enrichment bias. The main advantage of our method is that it relies on a sampling-based framework to perform gene-set enrichment analysis, which takes into consideration annotation uncertainties and avoids enrichment bias. In addition, gene-set enrichment results are often redundant and difficult to interpret, which we addressed by means of a semantic similarity-based algorithm to cluster similar terms and reduce the complexity of the enriched pathways.[47] We used such an approach to perform gene-set enrichment analysis in chapter 4 (in the context of AD) and chapter 5 (on longevity), and the resulting enriched pathways overlapped and improved those from previous studies. For example, in chapter 3 we studied five major pathways previously associated with AD from literature, while in chapter 4 we estimated these pathways using our sampling-based approach. As a result, in chapter 4 we clustered together all immune-related pathways (immune response and endocytosis), and added a cluster of pathways pointing to synaptic plasticity and remodeling. This refined and fully automated classification of each variant's effect on pathways may be used to further improve pathway-specific PRSs for patient stratification. Next to the functional interpretation of GWAS, to compare genetic association statistics across different traits can highlight a shared genetic basis of different traits. The state-of-the-art method to do so is LD-score regression and genetic correlation analysis, however, these approaches are limited to studies with large sample size.[48] An alternative approach is either to visually explore associ-

ation statistics in the same region across phenotypes or to browse existing datasets of variant-trait associations, such as the GWAS catalog.[49] We have addressed these limitations in *snpXplorer* (chapter 8), where we allow the visual superimposition of summary statistics from any study and the analysis of any given set of genetic variants in terms of functional enrichment and overlap with previous traits.

9.10 Becoming a cognitively healthy centenarian

In this thesis, we attempted to discover genetic signatures that are associated with becoming a cognitively healthy centenarian, which likely represents the ideal model of aging in good cognitive and physical conditions. Previous studies from our cohort showed that cognitively healthy centenarians were significantly more educated and performed significantly better in neuropsychological tests compared to centenarians from the same birth cohort.[50, 51] A possible explanation for this could be an advantageous genetic background as these individuals have a lower genetic risk to develop AD and other age-related diseases. Importantly, our studies shed light on the relationship between resilience against factors associated with AD risk, and longevity. Due to the difficulties in gathering a large number of cognitively healthy super-agers, this was never explored before and might explain why, apart from *APOE*, genetic variants influencing AD risk do not seem to influence longevity in large GWAS. In terms of functional implications, our findings from chapter 2, chapter 3, chapter 4, and chapter 5 pose great importance to genetic variants involved in immune-related pathways. These findings fit well in the context of a compromised immune response and a higher degree of chronic inflammation that are typically associated with aging. Genetic factors may result in improved regulatory mechanisms of the immune response, that might compensate for age-related changes and result in a better immune and metabolic system, at least in the context of Alzheimer's disease, that we observe in our cognitively healthy centenarians through the pathway-specific PRSs.[1, 13, 2] We also observed a similar pattern in the context of longevity, where we highlighted pathways such as differentiation processes, cellular response to stress, and nervous system development. The maintenance of these biological pathways may be associated with a slower progression of the aging mechanisms and with a concurrent delay of age-related diseases.

9.11 Drawbacks of studying centenarians

The study of individuals with extremely old age is not without flaws. Next to the difficulty in collecting a large sample of cognitively healthy centenarians due to their rareness in the population,[4] the definition of a cognitively healthy centenarian is not straightforward. In our study, all participants self-reported to be cognitively healthy, which was also confirmed by a family member or proxy-acquaintance. However, to date, there are no specific neuropsychological tests developed to score the cognitive performances of such old individuals. The 100-plus Study implements a battery of cognitive tests that typically are carried out in a memory clinic and, therefore, are specialized for the diagnosis of cognitive decline and dementia. While these tests allow to score different cognitive domains, they are not developed for extremely old individuals, and are usually not implemented in other centenarian studies. Our study has pioneered towards the identification of cognitive tests that may be more specific for centenarians.[51] Given this protocol, virtually any centenarian study around the world can identify a subset of individuals that maintained their cognitive and physical abilities in a same standardized way. This will eventually lead to more collaborative efforts, essential for genetic studies. An additional drawback relates to the short follow-up time available for these individuals, as well as difficulties in studying environmental factors. However, due to the high heritability estimates of longevity within families, a more feasible approach may be to investigate the children of the cognitively healthy centenarians, which should have inherited part of the protective genetic elements of their parents, are younger and thus could be followed-up for a longer time. This is an on-going effort in the 100-plus Study.

9.12 Future perspectives

The 100-plus Study is currently enrolling additional cognitively healthy centenarians and their family members, allowing the further exploration of the unique characteristics of these individuals. Our cohort of centenarians will continue to be part of large-collaborative studies investigating common (and rare) genetic factors associated with longevity and age-related diseases. Owing novel developments in the genetics field, the impact of larger structural genetic variations in diseases and longevity may be explored. In fact, with an increasing understanding of the architecture of our genome, it has become clear that structural variants (especilly those comprising repeated elements), are not only junk DNA, but may be implicated in diseases by interfering with

the normal gene transcription and translation processes. For AD, recently evidence of such mechanisms in the *ABCA7* gene was shown. The common *ABCA7* AD-associated variant is in linkage with an intronic variable number of tandem repeats (VNTR), a type of structural variation characterized by specific patterns of DNA that are repeated in tandem.[52] In this *ABCA7* VNTR, an excessive number of repeat units was associated with increased AD risk. To be able to infer the length of VNTRs at genomic level (either by direct measurement or by imputation), will allow to test whether such structural variations influence AD-risk.Apart from AD, such structural variations may also influence longevity directly, for example, the terminal parts of the chromosomes, *i.e.* the telomeres, are enriched with repetitive sequences, and the shortening of telomers is thought to be a direct consequence of aging. Therefore, the study of such variations will likely open new scenarios in the way we look at a genetic predisposition to aging and diseases. This will eventually allow to fill-in the missing heritability gap that still underlies the genetics of AD, longevity, and other age-relted traits. Finally, these effort will improve the predictive power of the PRS, which will hopefully be used as a valuable clinical parameter.

9.13 Conclusions

The main finding of this thesis was the characterization of the role of genetic variants associated with AD in cognitively healthy agers, and to discover molecular pathways that associate with the resilience against AD and other age-related diseases. We provided evidence for the effectiveness of using extreme phenotypes in genetic studies, the use of a PRS in a clinical setting for AD diagnosis, and actively improved the knowledge of the genetic factors that are associated with AD and longevity. The findings in this thesis are instrumental for future studies dealing with longevity, AD and other age-related disorders, and should inspire collaborative efforts especially among centenarian studies.

References

[1] Linda Partridge, Joris Deelen, and P. Eline Slagboom. "Facing up to the global challenges of ageing". In: *Nature* 561.7721 (Sept. 2018), pp. 45–56. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-018-0457-8.

[2] Angela R. Brooks-Wilson. "Genetics of healthy aging and longevity". In: *Human Genetics* 132.12 (Dec. 2013), pp. 1323–1338. ISSN: 1432-1203. DOI: 10.1007/s00439-013-1342-z.

[3] Thomas Perls. "Dementia-free centenarians". In: *Experimental Gerontology* 39.11 (Nov. 2004), pp. 1587–1593. ISSN: 05315565. DOI: 10.1016/j.exger.2004.08.015.

[4] Henne Holstege et al. "The 100-plus Study of Dutch cognitively healthy centenarians: rationale, design and cohort description". In: (Apr. 2018). DOI: 10.1101/295287.

[5] Robert Dantzer et al. "Resilience and immunity". In: *Brain, Behavior, and Immunity* 74 (Nov. 2018), pp. 28–42. ISSN: 08891591. DOI: 10.1016/j.bbi.2018.08.010.

[6] Erin J. Aiello Bowles et al. "Cognitive Resilience to Alzheimer's Disease Pathology in the Human Brain". In: *Journal of Alzheimer's Disease* 68.3 (Apr. 8, 2019). Ed. by Ozioma Okonkwo, pp. 1071–1083. ISSN: 13872877, 18758908. DOI: 10.3233/JAD-180942.

[7] Niccolò Tesi et al. "Centenarian controls increase variant effect sizes by an average twofold in an extreme case–extreme control analysis of Alzheimer's disease". In: *European Journal of Human Genetics* (Sept. 2018). ISSN: 1018-4813, 1476-5438. DOI: 10.1038/s41431-018-0273-5.

[8] Niccolò Tesi et al. "Immune response and endocytosis pathways are associated with the resilience against Alzheimer's disease". In: *Translational Psychiatry* 10.1 (Dec. 2020), p. 332. ISSN: 2158-3188. DOI: 10.1038/s41398-020-01018-7.

[9] A. M. Saunders et al. "Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease". In: *Neurology* 43.8 (Aug. 1993), pp. 1467–1472. ISSN: 0028-3878.

[10] W. J. Strittmatter et al. "Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease". In: *Proceedings of the National Academy of Sciences of the United States of America* 90.5 (Mar. 1993), pp. 1977–1981. ISSN: 0027-8424.

[11] E. H. Corder et al. "Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families". In: *Science (New York, N.Y.)* 261.5123 (Aug. 1993), pp. 921–923. ISSN: 0036-8075.

[12] E. H. Corder et al. "Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease". In: *Nature Genetics* 7.2 (June 1994), pp. 180–184. ISSN: 1061-4036. DOI: 10.1038/ng0694-180.

[13] David Melzer, Luke C. Pilling, and Luigi Ferrucci. "The genetics of human ageing". In: *Nature Reviews Genetics* (Nov. 2019). ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-019-0183-6.

[14] Joris Deelen et al. "A meta-analysis of genome-wide association studies identifies multiple longevity genes". In: *Nature Communications* 10.1 (Dec. 2019). ISSN: 2041-1723. DOI: 10.1038/s41467-019-11558-2.

[15] Jean-Charles Lambert et al. "Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease". In: *Nature Genetics* 41.10 (Oct. 2009), pp. 1094–1099. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.439.

[16] DESGESCO (Dementia Genetics Spanish Consortium), EADB (Alzheimer Disease European DNA biobank) et al. "A nonsynonymous mutation in PLCG2 reduces the risk of Alzheimer's disease, dementia with Lewy bodies and frontotemporal dementia, and increases the likelihood of longevity". In: *Acta Neuropathologica* (May 2019). ISSN: 0001-6322, 1432-0533. DOI: 10.1007/s00401-019-02026-8.

[17] William J. Astle et al. "The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease". In: *Cell* 167.5 (2016), 1415–1429.e19. ISSN: 1097-4172. DOI: 10.1016/j.cell.2016.10.042.

[18] Gleb Kichaev et al. "Leveraging Polygenic Functional Enrichment to Improve GWAS Power". In: *American Journal of Human Genetics* 104.1 (2019), pp. 65–75. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2018.11.008.

[19] Thorunn A. Olafsdottir et al. "Eighty-eight variants highlight the role of T cell regulation and airway remodeling in asthma pathogenesis". In: *Nature Communications* 11.1 (2020), p. 393. ISSN: 2041-1723. DOI: 10.1038/s41467-019-14144-8.

[20] Thomas J. Hoffmann et al. "A Large Multiethnic Genome-Wide Association Study of Adult Body Mass Index Identifies Novel Loci". In: *Genetics* 210.2 (2018), pp. 499–515. ISSN: 1943-2631. DOI: 10.1534/genetics.118.301479.

[21] Helen R. Warren et al. "Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk". In: *Nature Genetics* 49.3 (Mar. 2017), pp. 403–415. ISSN: 1546-1718. DOI: 10.1038/ng.3768.

[22] J. Nicholas Cochran et al. "The Alzheimer's disease risk factor CD2AP maintains blood–brain barrier integrity". In: *Human Molecular Genetics* 24.23 (Dec. 1, 2015), pp. 6667–6674. ISSN: 0964-6906, 1460-2083. DOI: 10.1093/hmg/ddv371.

[23] Hui Shi et al. "Genetic variants influencing human aging from late-onset Alzheimer's disease (LOAD) genome-wide association studies (GWAS)". In: *Neurobiology of Aging* 33.8 (Aug. 2012), 1849.e5–1849.e18. ISSN: 01974580. DOI: 10.1016/j.neurobiolaging.2012.02.014.

[24] Paul RHJ Timmers et al. "Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances". In: *eLife* 8 (Jan. 2019). ISSN: 2050-084X. DOI: 10.7554/eLife.39856.

[25] Aamira J. Huq et al. "Genetic resilience to Alzheimer's disease in *APOE* ϵ4 homozygotes: A systematic review". In: *Alzheimer's & Dementia* 15.12 (Dec. 2019), pp. 1612–1623. ISSN: 15525260. DOI: 10.1016/j.jalz.2019.05.011.

[26] Roberto Cabeza et al. "Maintenance, reserve and compensation: the cognitive neuroscience of healthy ageing". In: *Nature Reviews. Neuroscience* 19.11 (Nov. 2018), pp. 701–710. ISSN: 1471-0048. DOI: 10.1038/s41583-018-0068-2.

[27] Cristina Giuliani et al. "Centenarians as extreme phenotypes: An ecological perspective to get insight into the relationship between the genetics of longevity and age-associated diseases". In: *Mechanisms of Ageing and Development* 165 (July 2017), pp. 195–201. ISSN: 00476374. DOI: 10.1016/j.mad.2017.02.007.

[28] Paolo Garagnani et al. "Centenarians as super-controls to assess the biological relevance of genetic risk factors for common age-related diseases: a proof of principle on type 2 diabetes". In: *Aging* 5.5 (May 2013), pp. 373–385. ISSN: 1945-4589. DOI: 10.18632/aging.100562.

[29] Seunggeung Lee et al. "Rare-variant association analysis: study designs and statistical tests". In: *American Journal of Human Genetics* 95.1 (July 2014), pp. 5–23. ISSN: 1537-6605. DOI: 10.1016/j.ajhg.2014.06.009.

[30] Itziar de Rojas et al. *Common variants in Alzheimer's disease: Novel association of six genetic variants with AD and risk stratification by polygenic risk scores*. preprint. Genetic and Genomic Medicine, Nov. 2019. DOI: 10.1101/19012021.

[31] Céline Bellenguez et al. *New insights on the genetic etiology of Alzheimer's and related dementia*. en. preprint. Neurology, Oct. 2020. DOI: 10.1101/2020.10.01.20200659.

[32] Sultan Chaudhury et al. "Alzheimer's disease polygenic risk score as a predictor of conversion from mild-cognitive impairment". In: *Translational Psychiatry* 9.1 (Dec. 2019), p. 154. ISSN: 2158-3188. DOI: 10.1038/s41398-019-0485-7.

[33] Sultan Chaudhury et al. "Polygenic risk score in postmortem diagnosed sporadic early-onset Alzheimer's disease". In: *Neurobiology of Aging* 62 (Feb. 2018), 244.e1–244.e8. ISSN: 01974580. DOI: 10.1016/j.neurobiolaging.2017.09.035.

[34] "2012 Alzheimer's disease facts and figures". In: *Alzheimer's & Dementia* 8.2 (Mar. 2012), pp. 131–168. ISSN: 15525260. DOI: 10.1016/j.jalz.2012.02.001.

[35] Frank Dudbridge. "Power and Predictive Accuracy of Polygenic Risk Scores". In: *PLoS Genetics* 9.3 (Mar. 2013). Ed. by Naomi R. Wray, e1003348. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003348.

[36] J. Hardy. "The Amyloid Hypothesis of Alzheimer's Disease: Progress and Problems on the Road to Therapeutics". In: *Science* 297.5580 (July 19, 2002), pp. 353–356. ISSN: 00368075, 10959203. DOI: 10.1126/science.1072994.

[37] John Hardy. "Failures in Protein Clearance Partly Underlie Late Onset Neurodegenerative Diseases and Link Pathology to Genetic Risk". In: *Frontiers in Neuroscience* 13 (Dec. 5, 2019), p. 1304. ISSN: 1662-453X. DOI: 10.3389/fnins.2019.01304.

[38] Andrea B. Ganz et al. "Neuropathology and cognitive performance in self-reported cognitively healthy centenarians". In: *Acta Neuropathologica Communications* 6.1 (Dec. 2018), p. 64. ISSN: 2051-5960. DOI: 10.1186/s40478-018-0558-5.

[39] Sarah R. Elkin, Ashley M. Lakoduk, and Sandra L. Schmid. "Endocytic pathways and endosomal trafficking: a primer". In: *Wiener Medizinische Wochenschrift* 166.7 (May 2016), pp. 196–204. ISSN: 0043-5341, 1563-258X. DOI: 10.1007/s10354-016-0432-7.

[40] Scott A. Small and Gregory A. Petsko. "Endosomal recycling reconciles the Alzheimer's disease paradox". In: *Science Translational Medicine* 12.572 (Dec. 2, 2020), eabb1717. ISSN: 1946-6234, 1946-6242. DOI: 10.1126/scitranslmed.abb1717.

[41] Iris E. Jansen et al. "Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk". In: *Nature Genetics* 51.3 (Mar. 2019), pp. 404–413. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-018-0311-9.

[42] Alzheimer Disease Genetics Consortium (ADGC), et al. "Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing". In: *Nature Genetics* 51.3 (Mar. 2019), pp. 414–430. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-019-0358-2.

[43] Kyoko Watanabe et al. "Functional mapping and annotation of genetic associations with FUMA". In: *Nature Communications* 8.1 (Dec. 2017), p. 1826. ISSN: 2041-1723. DOI: 10.1038/s41467-017-01261-5.

[44] Christiaan A. de Leeuw et al. "MAGMA: Generalized Gene-Set Analysis of GWAS Data". In: *PLOS Computational Biology* 11.4 (Apr. 2015). Ed. by Hua Tang, e1004219. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004219.

[45] GTEx Consortium. "The Genotype-Tissue Expression (GTEx) project". In: *Nature Genetics* 45.6 (June 2013), pp. 580–585. ISSN: 1546-1718. DOI: 10.1038/ng.2653.

[46] Philipp Rentzsch et al. "CADD: predicting the deleteriousness of variants throughout the human genome". In: *Nucleic Acids Research* 47 (D1 Jan. 2019), pp. D886–D894. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gky1016.

[47] Fran Supek et al. "REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms". In: *PLoS ONE* 6.7 (July 2011). Ed. by Cynthia Gibas, e21800. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0021800.

[48] Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies". In: *Nature Genetics* 47.3 (Mar. 2015), pp. 291–295. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.3211.

[49] Annalisa Buniello et al. "The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019". In: *Nucleic Acids Research* 47 (D1 Jan. 2019), pp. D1005–D1012. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gky1120.

[50] Nina Beker et al. "Longitudinal Maintenance of Cognitive Health in Centenarians in the 100-plus Study". In: *JAMA Network Open* 3.2 (Feb. 2020), e200094. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2020.0094.

[51] Nina Beker et al. "Neuropsychological Test Performance of Cognitively Healthy Centenarians: Normative Data From the Dutch 100-Plus Study: COGNITIVE PERFORMANCE IN CENTENARIANS". In: *Journal of the American Geriatrics Society* 67.4 (Apr. 2019), pp. 759–767. ISSN: 00028614. DOI: 10.1111/jgs.15729.

[52] Arne De Roeck, Christine Van Broeckhoven, and Kristel Sleegers. "The role of ABCA7 in Alzheimer's disease: evidence from genomics, transcriptomics and methylomics". In: *Acta Neuropathologica* 138.2 (Aug. 2019), pp. 201–220. ISSN: 0001-6322, 1432-0533. DOI: 10.1007/s00401-019-01994-1.

10. Summary

10.1 English summary

One important accomplishment of humankind is the extension of the average life expectancy. However, a consequence of an aged population is the increased prevalence of age-related diseases, and, as a consequence, an increasing fraction of individuals will spend part of their old age in disability or dependence on others. Of all age-related diseases, a major contribution to poor health is cognitive decline and dementia, of which Alzheimer's disease (AD) is the most common type. However, dementia and AD are not inevitable: in fact, a small proportion of the population (<0.1%) reaches at least 100 years of age while maintaining a high level of cognitive and physical functions, so-called *cognitively healthy centenarians*. To investigate the genetic and environmental factors that characterize these individuals, the 100-plus Study was initiated.

Alzheimer's disease is a progressive neurodegenerative disorder characterized by loss of cognitive functions that leads to loss of independence, and death. Currently, there is no effective treatment to prevent of to slow down AD progression. The main risk factor for AD is aging: while the disease is rare before the age of 65 years, the prevalence increases exponentially and reaches ∼ 40% per year at 100 years of age. Next to aging, genetic factors play an important role as heritability estimates range 60-80%. The largest common genetic risk factor for AD is *APOE* genotype,which leads up to 30-fold increased risk for the disease. In addition to *APOE*, today ∼80 single nucleotide polymorphisms have been associated with the modulation of the risk of AD. Furthermore, these findings have been pivotal to understand the molecular events that are associated with AD development. While the most accepted hypothesis explaining AD pathogenesis puts amyloid accumulation at the basis of the molecular cascade that leads to cognitive decline, genetic studies have led to an evolution of this traditional hypothesis to encompass more complex aspects of the disease. As such, today it is thought that a dysregulation of the endo-lysosomal trafficking and immune systems are also major causal pathways of AD. Since AD is lethal at old age, it would be expected that genetic variants that increase the risk of AD, should affect negatively human lifespan.In fact, the main genetic risk factor for AD, *APOE* genotype, is also the largest genetic risk factor for longevity. Surprisingly, apart from *APOE*, none of the other genetic variants known to affect AD, also affects human longevity. Genetic factors are also though to affect human longevity: in fact, heritability of longevity up to ∼70 years of age ranges 10-25%, but to reach higher ages, we become increasingly dependent on the

favourable genetic elements in our genomes. In addition to *APOE*, several genetic variations have been found to affect human longevity, but these factors were not confirmed in different studies and populations, likely reflecting both technical and biological differences.Nevertheless, a common term to all previous genetic studies of longevity is that the genetic variants found to influence human longevity were previously associated with the risk of several age-related diseases. This suggests that an extended human lifespan is associated with a lower genetic risk of age-related diseases. Given this

background, the overall objective of this thesis was to investigate the genetic factors underlying extreme human longevity and the escape of Alzheimer's disease, for which we explore the genetic architecture of the *cognitively healthy centenarians* from the 100-plus Study.

10.1.1 Part I

In the first part of the thesis, which comprises **chapter 2-5**, we focus on the comparison of the cognitively healthy centenarians with young AD patients and population controls in the context of Alzheimer's disease and human longevity. In **chapter 2**, we exploited extreme phenotypes in the genetic

research of AD by comparing extreme controls, *i.e.* cognitively healthy centenarians, and extreme AD cases, *i.e.* relatively young AD patients, in a case-control study of AD. We report that cognitively healthy centenarians have a lower frequency of genetic variants associated with increased AD risk compared to the general population, and a higher frequency of protective genetic variants. This led to a 2-fold enrichment in the variant effect-size, showing that the use of extreme phenotypes in genetic studies of complex traits is profitable. In **chapter 3**, we investigated the molecular pathways

that are known to play a role in AD pathogenesis and their association with the resilience against AD. In this study, we combined the effect of multiple variants together into polygenic risk scores (PRS) and pathway-specific polygenic risk scores, which incorporate the effect of multiple genetic variants acting on the same molecular pathway. We report that cognitively healthy centenarians have the lowest PRS and pathway-specific PRS for all major AD-associated pathways. Moreover, only the PRS of immune system and endocytosis pathways significantly influenced the resilience against AD, even after excluding *APOE* variants. In **chapter 4**, we attempted

to disentangle the effect on longevity from the effect on AD risk of the genetic variants that are associated with AD. We found that most genetic variants that increase the risk of AD were associated with lower odds of longevity. Based on our analysis, most AD-associated variants negatively affect longevity through their increased risk of AD. However, a subset of variants preferentially involved in immune-related processes seemed to affect not only AD but also other age-related diseases, such that the cumulative effect on longevity was larger than the effect on AD alone. In **chapter**

5, we focused on human longevity and, using publicly available data, we constructed a polygenic risk score (PRS) that associated with becoming a cognitively healthy centenarians and independently with survival. This PRS included 330 genetic variants, did not include *APOE* variants, associated with up to 4-years longer survival, and showed functional enrichment for hallmarks of longevity, such as slow cell differentiation and replacement, and regulation of oxidative stress.

10.1.2 Part II

In the second part of the thesis, we present the contribution of the cognitively healthy centenarians from the 100-plus Study to large, collaborative GWAS of AD and longevity. In **chapter 6**, we combined clinical studies of

AD and by-proxy studies of AD into one of the largest GWAS of AD. This collaborative effort led to the discovery of six additional genetic variants associated with AD. Our findings reinforced the role of β-amyloid processing and immune response as central biological pathways in AD. Furthermore, we add on the growing literature showing the applicability of polygenic risk score (PRS) of AD in order to stratify patients based on their genetic background and to identify those at highest risk for the disease. In **chapter**

7, we collaborated on the, to date, largest GWAS of longevity. We introduced a new, unbiased, method to identify cases (*i.e.* the long-lived individuals) and controls, based on country- and sex-specific survival percentiles. In addition to *APOE* variants, we found a novel association near *GPR78* gene, and through genetic correlation and gene expression analyses, we showed a marked overlap between the genetics of diseases and the genetics of longevity. In **chapter 8**, the last chapter, we present *snpXplorer*, a tool freely available

to the scientific community to explore summary statistics of genetic stud-

ies, compare levels of association between different traits, and functionally annotate sets of genetic variants.

10.2 Nederlandse samenvatting

De verlenging van de gemiddelde levensverwachting is een belangrijke prestatie van de mensheid, maar heeft wel de vergrijzing van de bevolking als gevolg. Hierdoor komen leeftijdsgebonden ziekten steeds vaker voor en is een steeds groter deel van de mensen op hun oude dag invalide of afhankelijk van anderen. Van alle leeftijdsgebonden ziekten dragen cognitieve achteruitgang en dementie, waarvan de ziekte van Alzheimer de meest voorkomende is, in belangrijke mate bij aan een slechte gezondheid. Dementie en AD zijn echter niet onvermijdelijk: een klein deel van de bevolking (<0,1%) bereikt ten minste de leeftijd van 100 jaar en behoudt daarbij een hoog niveau van cognitieve en fysieke functies; de zogenaamde *cognitief gezonde honderdjarigen*. Om de genetische en omgevingsfactoren te onderzoeken die deze individuen maken tot wie ze zijn, werd het 100-plus Onderzoek opgezet.

De ziekte van Alzheimer is een progressieve neurodegeneratieve aandoening die wordt gekenmerkt door verlies van cognitieve functies met als gevolg verlies van onafhankelijkheid en uiteindelijk de dood. Momenteel is er geen effectieve behandeling om de progressie van de ziekte van Alzheimer te voorkomen of te vertragen. De belangrijkste risicofactor voor de ziekte van Alzheimer is veroudering: hoewel de ziekte zeldzaam is voor de leeftijd van 65 jaar, neemt de prevalentie exponentieel toe en bereikt ~40% per jaar op de leeftijd van 100 jaar. Naast veroudering spelen genetische factoren een belangrijke rol, waarbij de erfelijkheidsschattingen variëren van 60-80%. De meest voorkomende genetische risicofactor voor de ziekte van Alzheimer is het *APOE*-genotype, dat leidt tot een tot 30-voudig verhoogd risico op de ziekte. Naast *APOE* zijn vandaag ~80 single-nucleotide polymorfismen in verband gebracht met de modulatie van het risico op de ziekte van Alzheimer. Bovendien zijn deze bevindingen van cruciaal belang geweest om de moleculaire processen te begrijpen die in verband worden gebracht met de ontwikkeling van de ziekte van Alzheimer. De meest geaccepteerde hypothese over de pathogenese van de ziekte van Alzheimer stelt dat de accumulatie van amyloïd het begin is van een moleculaire cascade die leidt tot cognitieve achteruitgang. Genetische studies hebben geleid tot een herziening van deze traditionele hypothese die nu ook de complexe aspecten van de ziekte omvat. Zo wordt tegenwoordig gedacht dat een ontregeling van de endo-lysosomale processen in de cel en het immuunsysteem belangrijke causale routes van de ziekte van Alzheimer zijn. Aangezien de ziekte van Alzheimer dodelijk is op oudere leeftijd, zou men verwachten dat genetische varianten die het risico op de ziekte van Alzheimer verhogen, een negatieve invloed op de levensduur

van mensen hebben. Dit is ook zo want de belangrijkste genetische risicofactor voor de ziekte van Alzheimer, het APOE-genotype, is ook de grootste genetische risicofactor voor levensduur. Verrassend genoeg heeft, afgezien van *APOE*, geen van de andere genetische varianten waarvan bekend is dat ze invloed hebben op de ziekte van Alzheimer ook invloed op levensduur. Levensduur is echter wel erfelijk bepaald; de erfelijkheid van de levensduur tot ~70 jaar wordt geschat op 10-25%, maar om hogere leeftijden te bereiken, worden we steeds afhankelijker van de gunstige genetische elementen in ons genoom. Naast *APOE* zijn verschillende genetische variaties gevonden die de menselijke levensduur beïnvloeden, maar deze factoren werden niet bevestigd in verschillende studies en populaties, waarschijnlijk als gevolg van zowel technische als biologische verschillen. Niettemin is een gemeenschappelijk kenmerk van alle eerdere genetische studies van levensduur dat de genetische varianten die de levensduur van de mens bleken te beïnvloeden, eerder in verband werden gebracht met het risico van verschillende leeftijdsgebonden ziekten. Dit suggereert dat een langere levensduur van de mens geassocieerd is met een lager genetisch risico op leeftijdsgebonden ziekten.

Gezien deze achtergrond was het algemene doel van dit proefschrift om de genetische factoren te onderzoeken die ten grondslag liggen aan extreme menselijke levensduur en het ontsnappen aan de ziekte van Alzheimer. Hiervoor onderzochten we de genetische architectuur van patiënten met de ziekte van Alzheimer en de cognitief gezonde honderdjarigen uit de 100-plus Studie.

10.2.1 Part I

In het eerste deel van het proefschrift (**hoofdstukken 2-5**) richtten we ons op de verschillen in erfelijke factoren tussen cognitief gezonde honderdjarigen, jonge patiënten met de ziekte van Alzheimer en een controle groep uit de algemene bevolking.

In **hoofdstuk 2** hebben we gekeken naar de toegevoegde waarde van de analyse van extreme fenotypes in het genetisch onderzoek naar de ziekte van Alzheimer. We vergeleken extreme controles, d.w.z. cognitief gezonde honderdjarigen, met extreme jonge patiënten met de ziekte van Alzheimer. Wij vonden dat cognitief gezonde honderdjarigen een lagere frequentie van genetische risico varianten hebben en een hogere frequentie van beschermende genetische varianten. Gemiddeld zagen we een 2-voudige verrijking in de variant effect-grootte, waaruit blijkt dat het gebruik van extreme feno-

types kunnen helpen bij genetische studies naar de oorsprong van de ziekte van Alzheimer.

In **hoofdstuk 3** onderzochten we de moleculaire pathways waarvan bekend is dat ze een rol spelen in de pathogenese van de ziekte van Alzheimer en hun associatie met de weerbaarheid tegen de ziekte van Alzheimer. In deze studie hebben we het effect van meerdere varianten samengevoegd tot polygene risicoscores (PRS) en pathway-specifieke polygene risicoscores. Deze laatste omvatten het effect van meerdere genetische varianten die inhaken op dezelfde moleculaire pathway. Wij vonden dat cognitief gezonde honderdjarigen de laagste PRS en pathway-specifieke PRS hebben voor alle belangrijke Alzheimer-geassocieerde pathways. Bovendien bleken alleen de PRS van de immuunsysteem- en endocytose-pathways een significante invloed te hebben op de weerbaarheid tegen de ziekte van Alzheimer, zelfs na het uitsluiten van *APOE*-varianten.

In **hoofdstuk 4** hebben we geprobeerd om het effect van genetische varianten op de levensduur los te koppelen van het effect op het risico op de ziekte van Alzheimer. We ontdekten dat de meeste genetische varianten die het risico op de ziekte van Alzheimer verhogen, geassocieerd waren met een lagere kans op een lange levensduur. Op basis van onze analyse hebben de meeste Alzheimer-geassocieerde varianten een negatieve invloed op de levensduur door hun verhoogde risico op de ziekte van Alzheimer. Echter, een subset van varianten die bij voorkeur betrokken zijn bij immuungerelateerde processen leken niet alleen invloed te hebben op de ziekte van Alzheimer, maar ook op andere ouderdomsziekten, zodat het cumulatieve effect op de levensduur groter was dan het effect op de ziekte van Alzheimer alleen.

In **hoofdstuk 5** richtten we ons op de menselijke levensduur en maakten we gebruik van publiek beschikbare gegevens om een polygene risicoscore te maken die de kans verhoogd om cognitief gezond honderd jaar te worden en ook associeerde met overleving. Deze polygene risicoscore bevatte 330 genetische varianten (geen *APOE*-varianten) en associeerde tot 4 jaar langere overleving. De varianten in de polygene risicoscore toonde functionele verrijking voor kenmerken van een lange levensduur, zoals langzame celdifferentiatie en vervanging, en regulatie van oxidatieve stress.

10.2.2 Part II

In **hoofdstuk 6** hebben we klinische studies van de ziekte van Alzheimer en by-proxy studies van de ziekte van Alzheimer gecombineerd in één van

de grootste GWAS. Deze gezamenlijke inspanning leidde tot de ontdekking van zes extra genetische varianten die geassocieerd zijn met de ziekte van Alzheimer. Onze bevindingen versterken de rol van beta-amyloïd verwerking en immuunrespons als centrale biologische pathways in de ziekte van Alzheimer. Verder dragen we bij aan de groeiende literatuur die de toepasbaarheid van polygene risicoscore van AD aantoont. Met de polygene risicoscore konden we patiënten stratificeren op basis van hun genetische achtergrond en diegenen met het hoogste risico op de ziekte identificeren.

In **hoofdstuk 7** werkten we mee aan de, tot nu toe, grootste GWAS van langlevendheid. We introduceerden een nieuwe, onbevooroordeelde methode om gevallen (d.w.z. de langlevende individuen) en controles te identificeren, gebaseerd op land- en geslachtsspecifieke overlevingspercentielen. Naast *APOE*-varianten vonden we een nieuwe associatie in de buurt van het GPR78-gen, en door middel van genetische correlatie en genexpressieanalyses toonden we een duidelijke overlap aan tussen de genetica van ziekten en de genetica van een lang leven.

In **hoofdstuk 8**, het laatste hoofdstuk, presenteren we *snpXplorer*, een instrument dat vrij beschikbaar is voor de wetenschappelijke gemeenschap om samenvattende statistieken van genetische studies te verkennen. *snpXplorer* combineert meerdere niveaus van associatie en verricht functionele annotatie van sets van genetische varianten.

10.3 Riassunto in Italiano

Un importante traguardo del genere umano consiste nell'aumento dell'aspettativa di vita media. Tuttavia, una consequenza di una crescente popolazione anziana risulta essere l'aumento di diverse patologie legate all'invecchiamento. Ne consegue che una frazione in aumento di individui spenderà parte della loro età avanzata in disabilità e/o dipendenza da altri. Tra tutte le patologie collegate all'invecchiamento, il declino cognitivo e la demenza, di cui la malattia di Alzheimer è la causa più frequente, rappresentano le cause maggiori di impedimento. Ciò nonostante, lo sviluppo di demenza e/o malattia di Alzheimer non è una conseguenza inevitabile dell'invecchiamento: infatti, una piccola frazione di individui (<0.1%) raggiunge età superiori ai 100 anni mantenendo un livello sorprendentemente alto di funzioni cognitive e fisiche, i cosiddetti *centenari cognitivamente sani*. Per individuare fattori genetici ed ambientali che caratterizzano questi speciali individui, lo studio 100-plus è stato iniziato.

La malattia di Alzheimer è una patologia caratterizzata da un progressivo deterioramento delle funzioni cognitive che porta a mancanza di indipendenza, risultando letale. Attualmente, non ci sono rimedi e/o trattamenti farmacologici in grado di prevenire, attenuare, o revertire il progresso della malattia. Il fattore di rischio maggiore per lo sviluppo della malattia è l'età: mentre la patologia risulta essere rara prima dei 65 anni, la prevalenza aumenta esponenzialmente all'aumentare dell'età, e raggiunge il ∼40% all'anno a 100 anni di età. Inoltre, fattori genetici giocano un ruolo centrale nello sviluppo della malattia dato che l'ereditarietà della malattia di Alzheimer varia tra 60% e 80%. Il maggiore fattore di rischio genetico per la malattia di Alzheimer è dato dal genotipo del gene *APOE*. Il genotipo di *APOE* è determinato da due mutazioni genetiche a livello del gene *APOE*. Nella sua forma a più alto rischio, il genotipo di *APOE* aumenta il rischio di sviluppare la malattia di Alzheimer fino a 30 volte. Oltre ad *APOE*, oggi conosciamo ∼80 singole mutazioni genetiche che influenzano significativamente il rischio di sviluppare la patologia. La scoperta di questi fattori di rischio ha permesso l'identificazione dei processi molecolari che sono associati allo sviluppo della malattia. L'ipotesi più accreditata per lo sviluppo della malattia di Alzheimer pone l'accumulo di frammenti della proteina amiloide nel cervello come evento scatenante la cascata molecolare che porta al declino cognitivo. Tuttavia, studi genetici hanno evidenziato l'importanza di altri processi molecolari. Di conseguenza, si è assistito ad un'evoluzione della tradizionale ipotesi amiloidea in modo da includere aspetti più complessi della malattia.

Oggi, si pensa che una deregolazione dei sistemi endo-lisosomiali e del sistema immunitaria siano a loro volta processi centrali per lo sviluppo della malattia di Alzheimer.

Dato che la malattia di Alzheimer risulta letale ad età avanzata, ci si aspetterebbe che mutazioni genetiche che aumentano il rischio di sviluppare la malattia, abbiano un effetto negativo su longevità e sopravvivenza. Infatti, il maggiore fattore di rischio genetico per la malattia di Alzheimer, il genotipo di *APOE*, rappresenta anche il principale fattore genetico che influenza la longevità. Specificatamente, i genotipi di *APOE* che aumentano il rischio di sviluppare la malattia di Alzheimer, sono anche associati a una ridotta longevità. Soprendentemente, ad esclusione di *APOE*, nessuna della altre mutazioni genetiche associate alla malattia di Alzheimer, influenzano significativamente la longevità umana.

In precedenza, numerosi studi hanno esaminato il contributo dei fattori genetici nel modificate la durata della vita umana. Ne emerge un panorama contrastante: infatti l'ereditarietà della longevità fino a ∼70 anni risulta essere relativamente bassa (10-25%), tuttavia, per raggiungere età più avanzate, diventiamo sempre più dipendenti dai fattori favorevoli nascosti nel nostro genoma. In altre parole, più si invecchia, più i fattori genetici diventano importanti. Oltre ad *APOE*, diverse altre mutazioni genetiche sono state associate a longevità, anche se l'effetto di queste mutazioni genetiche non è stato confermato in diversi studi, od in diverse popolazioni. Queste divergenze tra studi probabilmente riflettono sia problematiche tecniche (di set-up dello studio e/o di metodologie statistiche), sia differenze a livello biologico. Ciò nonostante, un fattore comune a tutti i precedenti studi di genetica di longevità è che le mutazioni genetiche identificate erano state precedentemente associate a diverse patologie conseguenti l'invecchiamento. Questo suggerisce che una durata più lunga della vita umana dipende da una predisposizione genetica che diminuisce il rischio di sviluppare patologie associate all'età avanzata.

Nel complesso, lo scopo di questa tesi consiste nello studio dei fattori genetici alla base dell'estrema longevità e dalla resilienza nei confronti della malattia di Alzheimer che osserviamo nei centenari cognitivamente sani dello Studio 100-plus.

10.3.1 Prima parte

La prima parte di questa tesi, corrispondente ai **capitoli 2-5**, focalizza sulla comparazione tra centenari cognitivamente sani, pazienti affetti da Alzheimer

ed individui adulti sani nella popolazione (controlli), nel contesto dei fattori genetici associati alla malattia di Alzheimer, e longevità.

Nel **capitolo 2**, abbiamo utilizzato fenotipi estremi nella ricerca genetica applicata alla malattia di Alzheimer. Abbiamo contrapposto controlli estremi, *i.e.* centenari cognitivamente sani, a casi estremi, *i.e.* pazienti affetti da Alzheimer con un'età relativamente bassa.Abbiamo trovato che i centenari avevano una frequenza più bassa di mutazioni genetiche associate ad un aumento del rischio di Alzheimer (rispetto alla popolazione generale), ed una frequenza più alta di mutazioni genetiche che risultano protettive nei confronti della malattia (rispetto alla popolazione generale). Una consequenza pratica di questo studio è che l'utilizzo di fenotipi estremi nella ricerca di fattori genetici associati a patologie complesse (come la malattia di Alzheimer), è redditizio.

Nel **capitolo 3**, abbiamo studiato i processi molecolari che svolgono un ruole importante nello sviluppo e nel processo di resilienza contro la malattia di Alzheimer. In questo studio, abbiamo combinato l'effetto di multiple mutazioni genetiche in un valore di rischio poligenico (polygenic risk score, PRS). Questi valori di rischio poligenici quantificano il rischio genetico di sviluppare una determinata patologia, in questo caso la malattia di Alzheimer. Di conseguenza, più il valore di rischio è alto, maggiore è il rischio di sviluppare la patologia. Inoltre, abbiamo combinato l'effetto di più mutazioni che agiscono a livello del medesimo processo molecolare, creando un valore di rischio poligenico specifico per ciascun processo molecolare. Abbiamo trovato che i centenari possedevano i valori di rischio poligenici (PRS) più bassi in una comparazione tra centenari, pazienti affetti da Alzheimer, ed individui adulti sani nella popolazione. Soprattutto, abbiamo identificato che i valori di rischio poligenici specifico per il sistema immunitario ed il sistema endosomiale erano associati significativamente alla resilienza contro la malattia di Alzheimer, anche escludendo il fattore genetico di *APOE*.

Nel **capitolo 4**, abbiamo ragionato che mutazioni genetiche che aumentano il rischio di sviluppare la malattia di Alzheimer dovrebbero essere associate ad una maggiore mortalità, e di conseguenza, ad un rischio minore di longevità. In questo studio, abbiamo tentato di separare l'effetto su longevità da quello su Alzheimer da parte delle mutazioni genetiche che sono associate al rischio di sviluppare la malattia di Alzheimer. Abbiamo trovato che la maggioranza delle mutazioni genetiche che aumentano il rischio di sviluppare Alzheimer sono anche associate a minori probabilità di longevità. In base alla nostra analisi, la maggioranza delle mutazioni genetiche diminuisce la longevità a causa del rischio aumentato di Alzheimer. Tuttavia, un sot-

togruppo di mutazioni specificatamente coinvolte in processi immunitari, conferisce protezione non solo nei confronti della malattia di Alzheimer, ma anche nei confronti di altre patologie dell'età avanzata, cosicchè l'effetto cumulativo sulla longevità risulti maggiore dell'effetto su Alzheimer da solo.

Nel **capitolo 5** abbiamo focalizzato sulla longevità umana, ed utilizzando dati pubblicamente disponibili, abbiamo costruito un valore di rischio poligenico (PRS) che associava significativamente con il diventare un centenario, ed in un campione independente di invididui di mezza età, con sopravvivenza. Questo PRS includeva 330 mutazioni genetiche comuni nella popolazione generale, non includeva *APOE*, e determinava fino a 4 anni in più di sopravvivenza. A livello molecolare, queste mutazioni hanno effetto a livello di processi che solitamente sono alterati in età avanzata, come una ridotta velocità di differenziazione e sostituzione cellulare, e la regolazione dello stress ossidativo.

10.3.2 Seconda parte

Nella seconda parte della tesi,abbiamo presentato il contributo dei centenari dello Studio 100-plus in grandi studi di associazione genomica (GWAS) di Alzheimer e longevità.

Nel **capitolo 6**, abbiamo combinato studi clinici di Alzheimer e studi di proxy (o di delega) di Alzheimer in uno dei piu grandi GWAS di Alzheimer. Questo studio, che includeva più di mezzo milione di individui in totale, ha portato alla scoperta di 6 nuove mutazioni genetiche associate ad Alzheimer. Questi risultati hanno rinforzato il ruolo del metabolismo dell'amiloide e del sistema immunitario come processi centrali nello sviluppo di Alzheimer. Inoltre, abbiamo sviluppato un valore di rischio poligenico (PRS) significativamente predittivo per la malattia di Alzheimer, che potrebbe essere utilizzato nella clinica per la stratificazione di pazienti di Alzheimer e l'identificazione di individui con il rischio più alto di sviluppare la malattia.

Nel **capitolo 7**, abbiamo partecipato al più grande (fino ad ora) studio di associazione genomica (GWAS) di longevità. In questo studio, abbiamo introdotto un metodo nuovo ed imparziale per l'idenficazione di casi (individui longevi) e controlli, basato su percentili di sopravvivenza specifici per paese e sesso. Oltre ad *APOE*, abbiamo individuato una nuova mutazione vicino al gene *GPR78*, ed attraverso analisi di correlazione genetica ed espressione genica, abbiamo identificato una soprapposizione tra la genetica della longevità e quella di alcune patologie dell'età avanzata.

Nel **capitolo 8**, l'ultimo della tesi, presentiamo *snpXplorer*, uno stru-

mento gratuitamente disponibile ai ricercatori per esplorare associazioni derivanti da studi genetici, per comparare livelli di associazione genetica in diversi fenotipi o patologie, e per eseguire l'annotazione funzionale di qualsiasi gruppo di mutazioni genetiche.

11. Addendum

11.1 Acknowledgements

This thesis is is the result of the last 5 years of life, and without any doubt I would not have achieved this accomplishment without the contribution of many people, both in the working environment and also outside of it. Looking back at my *old me* now, i realise how much this process has modeled me as a scientist, and more importantly as a man. For this reason, I would like to take few pages to acknowledge the persons that I felt as part of this journey.

Dr. Holstege, Beste **Henne**, you were the first person I spoke with during my interviews, the first person who probably believed in me and saw something in me. Thank you for everything you have done for me, for having put amazing people around me and supervising me, and thanks for the many good advices you gave. It has been a real pleasure being part of the 100-plus Study. I am really proud of the group, how it has grown in the last 5 years, and the attention that is increasingly getting. All this was certainly because of your tenacity, and you are one of the main example, for me, that if you really believe in something, you should put all yourselft in it, and results will come. Thanks your being always yourself.

Professor Reinders, Beste **Marcel**, next to Henne, you have been one of the main person that guided me and helped me a lot during my PhD. You always have had so good suggestions, foods for thoughs and discussions, and every week I was looking forward to our meeting. I jelously still keep the ridiculously huge presentation of 500-600 slides that I was used to discuss with you. You introduced me to the Delft's group of PhDs, always had nice words if something unexpected happens, or to cheer me up after a rejection. One of the paper I am mostly attached to is the *snpXplorer*, and I could never forget your reaction when the paper was rejected at *Bioinformatics*, which gave me new energy to believe in what I was doing, and eventually we improved the manuscript, and managed to published in *Nucleic Acid Research*. I truly believe that I wouldn't have achieved this without your precious help. Thanks for being a superhero.

Professor van der Flier, Beste **Wiesje**, I will never forget your motivation, spirit of collaboration and positivity. You have always found the time to help me, give very thoughful comments and suggestions, and during the though Corona times, I was always looking forward to your weekly update in which, despite the sad lockdown times, you always tried to make people feel less

lonely.

From my supervisors I have learned how important is to work hard, that if you work hard at some point you will be paid off, that collaborations and networking can make the difference, and that positivity is the key to accomplish something great, even in though and desperate times. Sincerely, thanks to all of you.

A big thank goes to my promotion committee: Professor *D. Posthuma*, Professor *K. Sleegers*, Professor *J. Hardy*, Professor *M. Verhage* and Dr. *A. Ruiz* for accepting to review my thesis.You all have my sincere admiration for what you have accomplished in your scientific career and really hope we could start a collaboration in the future.

Another big thank goes to one of my daily supervisors, **Marc**: the bioinformatics and statistical genius behind the 100-plus group. Thanks for all your precious feedbacks, and for having helped me throughout my PhD, I really appreciated this. I always had the feeling we had compatible characters, both quite, both always ready to help. But, you also have the tendency to like complex things (*oh, your brain...*). I wish you all the luck and best wishes for what the future will bring you.

Next to Marc, **Sven**, you have been a real revolution in my PhD (for the better!) You were freshly out from your PhD when you started to supervise me, and probably because of this, I felt you like an older brother, an enormous source of good advices, always ready to help, cheer me up, but also to scold me when was necessary. I will never forget the Fridays in the office, most of the time no-one was there, just me and you, and then at the sign of *Almost weekend*, music was starting. Kinda the same routine, some goold old classics like Dire Straits and Pink Floyd, and sometimes dance mode, with the captain Gigi Dag in console. Man, that was crazy. I am so thankful to you for everything you have done for me, and I really believe your contribution to this thesis is remarkable. I am sure your scientific future will be extremely bright.

Thanks to **Jasper**, former PhD student in the 100-plus Study group. You have been an example for me, and when I think about how a real scientist should be, I immediately think about you. Always skeptical, always eager of knowing things and understanding how things work, you were definitely ahead of your time when you did your PhD. Thanks for everything, and I

wish you a bright career.

Thanks to **Ramon**, former student in the 100-plus Study and perhaps the first dutch friend I had in Amsterdam. Thanks for having introduced me to the Dutch world, and for having tried to make me learn Dutch with the *sentence of the week*, the *zin van de wek*. That did not really work, but I will always remember you and the good times spent together.

Thanks to **Nina, Debbie, Andrea, Sterre, Esther**,the wonderwall ladies of the 100-plus Study. Many of you are not in the group anymore, but I will never forget the funny moments together, the retreats, the Fridays afternoon, the 100-plus days and the *fietschallenges*. I truly wish you all the best for your future plans and life. It's been a pleasure to meet and work with you.

Thanks to the Delft group, especially **Stavros, Tamim, Alex, Christine, Soufiane, Ahmed, Meng**, most of us started our PhDs approximately together, and most of us are ending it together. I have learned a lot from you guys, hardcore bionformaticians. Thanks for the good times together at the retreats, and the funny afternoon at the Sport Center playing basketball or football. I am sure each of you guys will have a fruitful career because you guys rock!

Thanks to all students that have been doing their internship during my journey, both at the 100-plus group and in Delft. **Nick, Anne-Fleur, Christa**, thanks for the nice moments together in the office. And a special thanks to the students I supervise, **Francesca** and **Gerard**, I am really thankful to you guys, I hope I have been a good supervisor for you and for sure I learned a lot from our nice and inspiring discussions.

Thanks to the newer members of the 100-plus Study, **Mayca, Kimberly, Susan, Alex, Linda** that do an impressive job with the centenarians. Over the years the group has grown a lot, and now you guys have to bring the 100-plus study to a new level. But, I don't have any doubts you will do that. I am sorry we mostly have seen each other from the screen of a laptop and we could not share many experiences as with the former colleagues, and I wish you all the luck for your future.

Thanks to everyone that has been, is or will be part of the 100-plus Study. Everyone should be very proud of being part of this family. Although I am not sure what the future will bring me and whether I will continue to work

in the 100-plus Study, I will forever keep the people and the good memories in my heart.

A special thanks goes to the centenarians. I have **never** seen such incredible persons in my life. I will always remember the joy of life in the eyes of these persons, despite all they have seen in their lifes. I am still not sure whether I really want to become a centenarian, but you guys really tried to convince me very hard.

Thanks to all members of the Clinical Genetics department, we did not have a very deep relationship, but I have always appreciated crossing the corridors with you, and smiling.

It is crazy how (almost) everything that happened during these 5 years is somehow related to the job. Many colleagues became friends, and in this regard a special thanks goes to the Italian community within the Alzheimer Center in Amsterdam. Grazie a **Silvia**, la prima persona italiana che ho conosciuto, un esempio di amore e passione enorme per il proprio lavoro, ed una carica incredibile. Purtroppo non siamo riusciti a concludere il nostro progetto progetto, ma lo rimpiangeranno! Grazie a **Daniele**,altro esempio di brillante scienziato ed amico. Grazie a **Timothy**, sempre sul pezzo, sempre pronto ad uscire, fare, conoscere, senza freni. Non penso di aver conosciuto una persona più attiva in questi anni in Olanda. Grazie a **Linda**, senza la quale, la maggioranza delle persone che mi sono state vicine durante questo periodo, non ci sarebbero state. Collante incredibile, persona meravigliosa, sempre pronta a far festa. Anche se ci siamo un pò allontanati con gli anni, ci tengo a dire che voi tutti occuperete un posto speciale nel mio cuore per tutto quello che avete ed abbiamo fatto insieme.

When you move to a new country, you often need to start from scratch. And, next to the colleagues, the people you live with become your new family. And since housing in Amsterdam is kinda crazy, I can say I have a large family. Grazie a **Simone** e **Francesca**, i miei primi coinquilini a *Dintelstraat*. Ci siamo divertiti un bel pò, purtroppo non è andata troppo bene con la casa ma è stata un'esperienza. Thanks to **Lucas** and **Hugo**, former housemates in *Van Heenvlietlaan*, we really had a lot of fun, concerts, nights, drinks, and mice in the house. Thanks for all good moments, after all, we managed well to not destroy the house. After 2 years of sharing houses with other folks, me and *Giulia* decided to move together, alone. That's when we met **Edwin** and **Nelleke**, by far the best landords anyone can hope to find. Thanks for

everything you have done for us, sometimes we felt like you were our Dutch parents. And, of course, thanks for inviting us at the most amazing wedding I have ever been to.

Uno prova a distaccarsi da ciò che conosce, a provare cose nuove, a conoscere persone nuove e culture nuove. Poi arriva la realizzazione che le persone con cui ti trovi meglio sono quelle più simili a te. Un grazie enorme va alla vera famiglia italiana ad Amsterdam. Ok, non è una famiglia pura italiana all 100%, ma lo scheletro è tricolore.

E quindi grazie a **Paolo**, pisano doc *deh*, italiano doc, una bella parte di Toscana ad Amsterdam.Paolo ha la straordinaria capacità di farti sentire meglio, di strapparti un sorriso o una risata, sempre. Ne abbiamo passate tante, potremmo fare un film *Appartamento a Parigi*, ne abbiamo bevute tante, grazie di tutto. Gracias to Andrea, paraninph with me for Giulia's defense, and Paolo's girlfriend. I wish you guys a lot of happiness, you are a great couple!

Grazie a **Luigi**, in teoria collega, ma ci siamo visti si e no 2 volte in ufficio quindi non sei un collega! Nonostante te la tiri un pò visto che sei di Roma, sei un grande Luì! Ci siamo conosciuti poco prima della pandemia ma ne abbiamo passate di cene, bevute e soprattutto schitarrate. *Heeey how are you doing Justin here*, e la pasta e fagioli, e l'aeroporto fantasma di Pisa, mi ricorderò per sempre di tante cose.

Thanks to the international part of our group: thanks **Jeff**, crazy american guy, we shared an interesting hobby together. ;) Grazie a **Fra**, un fiume in piena, una forza della natura, e una cuoca che...mammamia! Grazie per gli inviti, per le organizzazioni, per esserci stata per Giulia sempre, non sarebbe stato lo stesso senza di te.

Thanks to **Patrick**, the only Dutch guy in the group. We shared many epic moments, especially barcecues. Thanks for sharing important tricks of BBQinq, and inviting us to your place so often. Grazie a **Natascia**, un'altra forza della natura, sempre pronta a far festa ed a condividere una spettacolare carbonara.

Grazie ad **Erika**, compagna di Master a Bologna, compagna di dottorato in Olanda, persona con una tenacia, forza ed intelligenza incredibile. E da buona veneta, bevitrice numero 1. Grazie per tutte le serate e cene che

abbiamo condiviso in questi anni. Thanks to **Francisco**, Erika's boyfriend, for his amazing calm and good spanish vibe.

5 anni fa decisi di accettare la posizione di dottorato ad Amsterdam. 4 anni di contratto. Raramente si finisce entro 4 anni. Andare o non andare? *Andare*. Questa decisione è stata il risultato di come sono cresciuto, maturato, e delle mie esperienze. Mi sono sempre piaciute le esperienze all'estero, da solo e non, in cui ti devi arrangiare, fare, conoscere. Può andare male, ci possono essere eventi spiacevoli, ma quello che queste esperienze ti lasciano è sempre costruttivo. Tutte le decisioni che ho fatto sono state il risultato delle persone che ho incontrato, e alcune delle quali hanno avuto una grossa influenza su di me.

Grazie a **Luca** e **Alessandra**: ci siamo conosciuti alla triennale a Firenze, ma sono convinto che poche persone in vita mia mi abbiano influenzato così tanto. Non ci vediamo spesso, è vero, tutti da tre parti diverse d'Europa, ma vi penso spesso, e quanto sarebbe bello rivedersi, ritrovarsi, magari in *via della Palancola*. Grazie ad **Ale**, la persona che conosco fin da quando ho ricordi. Ne sono cambiate di cose, siamo cresciuti da quando giocavano a monopoli e mangiavamo caramelle gommose. Ti stimo molto per come ha preso in mano la tua vita. *For ever and ever*

Grazie agli amici di infanzia, quelli più stretti, grazie ad **Andrea**, **Raoul**, **Gianmarco**, **Simone**, **Lorenzo** e **Niccolò**. Le cose cambiano, ma alcune rimangono sempre uguali. Anche adesso, anche se torno 1-2 volte all'anno in Italia, è come non essersene mai andato, è come mettere in pausa e riavviare.

Un grande grazie a chi è venuto a trovarci ad Amsterdam. In 5 anni, sono felice di dire che ci sono state molte persone, ognuno portando qualcosa di speciale e lasciando un ricordo, una risata, un momento di felicità. E quindi grazie a **Raoul, Chiara, Nicola, Silvia, Matteo, Manolo, Francesca, Bruno, Francesca, Leonard, Silvia, Mjriam, Chiara, Gianmarco, Simone, Federico, Giulia, Gessica, Alessandro, Chiara**.

Grazie ad **Al** e **Mj**, che hanno allietato il mio periodo all'estero, e senza i quali non sarebbe stato lo stesso. Grazie a tante altre persone: grazie a chi c'è stato e chi ci sarà.

Grazie alla mia famiglia, che nonostante le mie scelte di *fuggire via*, mi hanno sempre supportato (e sopportato). E quindi grazie a mamma **Melina**,

babbo **Marco** e fratello **Federico**, grazie per i sacrifici che avete fatto e che mi hanno permesso di arrivare dove sono arrivato. Non ce l'avrei mai fatta senza il vostro costante aiuto e supporto, nei momenti felici e soprattutto in quelli bui. Una parte significativa di questo traguardo è senza ombra di dubbio vostra. Grazie anche a **Lorella, Alessandro,Charlie**, grazie agli zii, cugini ed ai nonni.

Un grazie speciale, in famiglia, è per te **Fede**, mio fratello. Grazie perchè ci sei sempre stato e sono sicuro ci sarai sempre. So che la mia partenza ha portanto un carico extra per te, mi dispiace e spero di poter ripagare. Grazie per tutte le volte che sei venuto a Pisa senza molto preavviso. Grazie per essere come sei. *Perchè comunque vada mio fratello ci sarà, grazie mamma, grazie pa...*

Cinque anni sono tanti, cinque anni lontano da casa possono essere pesanti, ti possono piegare. Ma non se hai accanto qualcuno con cui condividere gioie, dolori, pianti, lamenti, Netflix e risate. E quindi, *the last but not the least*, il più grande grazie va a te *Giulia*. Grazie per avermi spinto ad iniziare questo percorso, grazie per avermi sostenuto, per avermi seguito nonostante non ti andasse. Grazie per averci messa tutta te stessa, per aver iniziato un dottorato più per necessità che per scelta, grazie per tutto quello che abbiamo fatto insieme. Grazie per i viaggi, per le cene, per i bisticci sul sale e sull'olio, sulla cioccolata. Grazie per tutti i piccoli gesti che rendono migliori ogni giornata. Sono tanto orgoglioso di te, di tutto quello che hai fatto, e di quello che sei diventata. Questo capitolo insieme si avvicina alla fine, e non vedo l'ora di sapere quello che il futuro ci riserverà. Qualunque cosa, purchè insieme. Ancora non sappiamo dove, non sappiamo a fare cosa, sappiamo solo che saremo insieme, e in un posto soleggiato.

At the very end, I want to thank the Netherlands. It hasn't been easy all the time, especially from October to June, with the cold, the wind, the rain, the humidity (*oh my god where did i go?!*). Thanks to the Dutch because no matter what the wheather looks like, they do what their agenda tell them. Thanks to the Dutch for being inclusive and open-minded, and because they don't mind to switch to English if you are a lazy ignorant that did not want to learn the language (*yes, I am one of them*). Thanks to the Netherlands for being at the edge of international research, and for having allowed me to do my PhD here.

11.2 About the author

Niccolò was born in Florence on 19 May 1991, and grew in Quarrata, a small city in the beautiful tuscanian countryside. Nicco studied Medical Biotechnology in Florence for his bachelor, and then attended the International Master in Bioinformatics in Bologna, from which he graduated on September 2016. During his studies, he increased his international experiences with the Erasmus program, that let him spend 6 months in Barcelona (Spain) and Nijmegen (The Netherlands). In December 2016, he moved to The Netherlands for his PhD. Nicco is curious and always eager to learn something new. Genetics really fascinates him, it is amazing how a tiny molecule can influence the way we look like, behave, age and react to sicknesses. He is also very interested in the tech-world: computer, photography, videography, data analysis and machine learning. And because modern times can be stressful, he likes to escape this busyness by listening to some good-old records or going around with the skateboard.

11.3 Portfolio

Courses	**Year**
Doctoral Education Programme	2017
Functional Genomics and System Biology	2017
Presenting and Pitching training course	2020
Research Integrity	2020
International Conferences	**Year**
Poster presentation at Alzheimer Association International Conference (AAIC)	2018, 2021
Oral presentation at International Centenarian Conference (ICC)	2019,2021
Oral presentation at Alzheimer Association International Conference (AAIC)	2020
National Conferences	**Year**
Poster presentation at The Dutch Society for Research on Ageing (DuSRA)	2019
Poster presentation at BioSB	2017, 2018
Poster presentation at TN2 conference	2017
Annual meeting of Amsterdam Neuroscience	2018, 2019
Participation in VUmc Science Exchange Day (SED)	2018, 2019
Poster presentation at BioDay	2019
Oral presentation at BioSB	2020,2021
Other activities	**Year**
Supervision of 2 master thesis students at TUDelft	2020,2021
Organization of research group retreat	2019
Journal club at TUDelft	2019,2020,2021
Best presentation award at International Centenarian Conference (ICC)	2021

11.4 List of publications

In this thesis

Tesi, N., van der Lee, S.J., Hulsman, M. *et al.* Centenarian controls increase variant effect sizes by an average twofold in an extreme case–extreme control analysis of Alzheimer's disease. *Eur J Hum Genet* 27, 244–253 (2019). https://doi.org/10.1038/s41431-018-0273-5

Tesi, N., van der Lee, S.J., Hulsman, M. *et al.* Immune response and endocytosis pathways are associated with the resilience against Alzheimer's disease. *Transl Psychiatry* 10, 332 (2020). https://doi.org/10.1038/s41398-020-01018-7

Tesi, N., Hulsman, M., van der Lee, S.J. *et al.* The effect of Alzheimer's disease-associated genetic variants on longevity. *medRxiv* (2021). https://www.medrxiv.org/content/10.1101/2021.02.02.21250991v1.full.pdf+html

Tesi, N., van der Lee,S.J., Hulsman, M. *et al.* Polygenic Risk Score of Longevity Predicts Longer Survival Across an Age Continuum. *The Journals of Gerontology: Series A*, 76, 5, 750-759 (2021). https://doi.org/10.1093/gerona/glaa289

de Rojas, I., Moreno-Grau, S., **Tesi, N.**, *et al.* Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. *Nat Commun* 12, 3417 (2021). https://doi.org/10.1038/s41467-021-22491-8

Deelen, J., Evans, D.S., Arking, D.E. ,**Tesi, N.** *et al.* A meta-analysis of genome-wide association studies identifies multiple longevity genes. *Nat Commun* 10, 3669 (2019). https://doi.org/10.1038/s41467-019-11558-2

Tesi, N., van der Lee, S.J., Hulsman, M., Holstege, H. & Reinders, M.J.T *snpXplorer*: a web application to explore human SNP-associations and annotate SNP-sets. *Nucleic Acids Research*, 49,W1, W603-W612 (2021). https://doi.org/10.1093/nar/gkab410

Other publications

van der Lee, S.J., Conway, O.J., Jansen, I. *et al.* A nonsynonymous mutation in *PLCG2* reduces the risk of Alzheimer's disease, dementia with Lewy bodies and frontotemporal dementia, and increases the likelihood of longevity. *Acta Neuropathol* 138, 237–250 (2019). https://doi.org/10.1007/s00401-019-02026-8

Moreno-Martínez, D., Nomdedeu, M., Lara-Castillo, M.C., *et al.* XIAP inhibitors induce differentiation and impair clonogenic capacity of acute myeloid leukemia stem cells. *Oncotarget* 5(12):4337-4346 (2014). https://doi.org/10.18632/oncotarget.2016

Cornet-Masana, J.M., Moreno-Martínez, D., Lara-Castillo, M.C., *et al.* Emetine induces chemosensitivity and reduces clonogenicity of acute myeloid leukemia cells. *Oncotarget* 7(17):23239-23250 (2016). https://doi.org/10.18632/oncotarget.8096

Marneth, A., Prange, K., Al Hinai, A. *et al.* C-terminal *BRE* overexpression in 11q23-rearranged and t(8;16) acute myeloid leukemia is caused by intragenic transcription initiation. *Leukemia* 32, 828–836 (2018). https://doi.org/10.1038/leu.2017.280

Beker, N., Sikkes, S.A.M., Hulsman, M., *et al.* Longitudinal Maintenance of Cognitive Health in Centenarians in the 100-plus Study. *JAMA Netw Open* 3(2):e200094. (2020). https://doi.org/10.1001/jamanetworkopen.2020.0094

Bellenguez, C., Küçükali, F., Jansen, I., *et al.* New Insights on the Genetic Etiology of Alzheimer's and Related Dementia. *medRxiv* (2020). https://doi.org/10.1101/2020.10.01.20200659

11.5 List of theses from the Alzheimer Center Amsterdam

1. L. Gootjes: *Dichotic Listening, hemispheral connectivity and dementia* (14-09-2004)

2. K. van Dijk: *Peripheral Nerve Stimulation in Alzheimer's Disease* (16-01-2005)

3. R. Goekoop: *Functional MRI of cholinergic transmission* (16-01-2006)

4. R. Lazeron: *Cognitive aspects in Multiple Sclerosis* (03-07-2006)

5. N.S.M. Schoonenboom: *CSF markers in Dementia* (10-11-2006)

6. E.S.C. Korf: *Medial Temporal Lobe atrophy on MRI: risk factors and predictive value* (22-11-2006)

7. B. van Harten: *Aspects of subcortical vascular ischemic disease* (22-12-2006)

8. B. Jones: *Cingular cortex networks: role in learning and memory and Alzheimer's disease related changes* (23-03-2007)

9. L. van de Pol: *Hippocampal atrophy from aging to dementia: a clinical and radiological perspective* (11-05-2007)

10. Y.A.L. Pijnenburg: *Frontotemporal dementia: towards an earlier diagnosis* (05-07- 2007)

11. A.Bastos Leite:*Pathological ageing of the Brain* (16-11-2007)

12. E.C.W. van Straaten: *Vascular dementia* (11-01-2008)

13. R.L.C. Vogels: *Cognitive impairment in heart failure* (11-04-2008)

14. J. Damoiseaux: *The brain at rest* (20-05-2008)

15. G.B. Karas: *Computational neuro-anatomy* (19-06-2008)

16. F.H. Bouwman: *Biomarkers in dementia: longitudinal aspects* (20-06-2008)

17. A.A. Gouw: *Cerebral small vessel disease on MRI: clinical impact and underlying pathology* (20-03-2009)

18. H. van der Roest: *Care needs in dementia and interactive digital information provisioning* (12-10-2009)

19. C. Mulder: *CSF Biomarkers in Alzheimer's disease* (11-11-2009)

20. W. Henneman: *Advances in hippocampal atrophy measurement in dementia: beyond diagnostics* (27-11-2009)

21. S.S. Staekenborg: *From normal aging to dementia: risk factors and clinical findings in relation to vascular changes on brain MRI* (23-12-2009)

22. N. Tolboom: *Imaging Alzheimer's disease pathology in vivo: towards an early diagnosis* (12-02-2010)

23. E. Altena: *Mapping insomnia: brain structure, function and sleep intervention* (17-03- 2010)

24. N.A. Verwey: *Biochemical markers in dementia: from mice to men. A translational approach* (15-04-2010)

25. M.I. Kester: *Biomarkers for Alzheimer's pathology; Monitoring, predicting and understanding the disease* (14-01-2011)

26. J.D. Sluimer: *Longitudinal changes in the brain* (28-04-2011)

27. S.D Mulder: *Amyloid associated proteins in Alzheimer's Disease* (07-10-2011)

28. S.A.M. Sikkes: *Measuring IADL in dementia* (14-10-2011)

29. A. Schuitemaker: *Inflammation in Alzheimer's Disease: in vivo quantification* (27-01-2012)

30. K. Joling: *Depression and anxiety in family caregivers of persons with dementia* (02- 04- 2012)

31. W. de Haan: *In a network state of mind* (02-11-2012) (*Cum Laude*)

32. D. van Assema: *Blood-brain barrier P-glycoprotein function in ageing and Alzheimer's disease* (07-12-2012)

33. J.D.C. Goos: *Cerebral microbleeds: connecting the dots* (06-02-2013)

34. R. Ossenkoppele: *Alzheimer PEThology* (08-05-2013)

35. H.M. Jochemsen: *Brain under pressure: influences of blood pressure and angiotensinconverting enzyme on the brain* (04-10-2013)

36. A.E. van der Vlies: *Cognitive profiles in Alzheimer's disease: Recognizing its many faces* (27-11-2013)

37. I. van Rossum: *Diagnosis and prognosis of Alzheimer's disease in subjects with mild cognitive impairment* (28-11-2013)

38. E.I.S. Möst: *Circadian rhythm deterioration in early Alzheimer's disease and the preventative effect of light* (03-12-2013)

39. M.A.A. Binnewijzend: *Functional and perfusion MRI in dementia* (21-03-2014)

40. H. de Waal: *Understanding heterogeneity in Alzheimer's disease: A neurophysiological perspective* (25-04-2014)

41. W. Jongbloed: *Neurodegeneration: Biochemical signals from the brain* (08-05-2014)

42. E.L.G.E. Poortvliet-Koedam: *Early-onset dementia: Unraveling the clinical phenotypes* (28-05-2014)

43. A.C. van Harten: *The road less traveled: CSF biomarkers for Alzheimer's disease: Predicting earliest cognitive decline and exploring microRNA as a novel biomarker source* (07-02-2014)

44. A.M. Hooghiemstra: *Early-onset dementia: With exercise in mind* (03-12-2014)

45. L.L. Sandberg-Smits: *A cognitive perspective on clinical manifestations of Alzheimer's disease* (20-03-2015)

46. F.H. Duits: *Biomarkers for Alzheimer's disease, current practice and new perspectives* (01-04-2015)

47. S.M. Adriaanse: *Integrating functional and molecular imaging in Alzheimer's disease* (07-04-2015)

48. C. M ller: *Imaging patterns of tissue destruction - Towards a better discrimination of types of dementia* (01-05-2015)

49. M. del Campo Milàn: *Novel biochemical signatures of early stages of Alzheimer's disease* (19-06-2015)

50. M. R. Benedictus: *A vascular view on cognitive decline and dementia: relevance of cerebrovascular MRI markers in a memory clinic* (20-01-2016)

51. M. D. Zwan: *Visualizing Alzheimer's disease pathology. Implementation of amyloid PET in clinical practice* (03-03-2016)

52. E. Louwersheimer: *Alzheimer's disease: from phenotype to genotype* (21-06-2016)

53. W.A. Krudop: *The frontal lobe syndrome: a neuropsychiatric challenge* (23-09-2016)

54. E.G.B. Vijverberg: *The neuropsychiatry of behavioral variant frontotemporal dementia and primary psychiatric disorders: similarities and dissimilarities* (22- 09-2017)

55. F.T. Gossink: *Late Onset Behavioral Changes differentiating between bvFTD and psychiatric disorders in clinical practice* (20-04-2018)

56. M.A. Engels: *Neurophysiology of Dementia* (18-05-2018)

57. S.C.J. Verfaillie: *Neuroimaging in subjective cognitive decline: Incipient Alzheimer's disease unmasked* (12-09-2018)

58. M. ten Kate: *Neuroimaging in Predementia Alzheimer's Disease* (13-09-2018)

59. H.F.M. Rhodius-Meester: *Optimizing use of diagnostic tests in memory clinics: the next step* (24-09-2018)

60. E.A.J. Willemse: *Optimizing biomarkers in cerebrospinal fluid. How Laboratory reproducibility improves the diagnosis of Alzheimer's disease* (18-10-2018)

61. E. Konijnenberg: *Early amyloid pathology - Identical twins, two of a kind?* (25-06- 2019)

62. A.E. Leeuwis: *Connecting heart and brain; Vascular determinants of cognitive impairment and depressive symptoms* (02-07-2019)

63. J. Den Haan: *Imaging The Retina in Alzheimer's Disease* (12-09-2019)

64. A.C. van Loenhoud: *Cognitive reserve in Alzheimer's disease. A perspective on the flourishing and withering of the brain* (18-09-2019)

65. R.J. Jutten: *Capturing changes in cognition; Refining the measurement of clinical progression in Alzheimer's disease* (20-09-2019)

66. N. Legdeur: *Determinants of cognitive impairment in the oldest-old* (08-10-2019)

67. R. Slot: *Subjective cognitive decline-predictive value of biomarkers in the context of preclinical Alzheimer's disease* (14-11-2019)

68. N. Scheltens: *Understanding heterogeneity in Alzheimer's disease-a data driven approach* (17-12-2019)

69. L. Vermunt: *Secondary Prevention for Alzheimer Disease - Timing, Se-*

lection and Endpoint of Clinical Trials (13-03-2020)

70. L.M.P. Wesselman: *Lifestyle and brain health - exploring possibilities of an online intervention in non-demented elderly* (01-04-2020)

71. I.S. van Maurik: *Interpreting biomarker results in patients with mild cognitive impairment to estimate prognosis and optimize decision making* (12-05-2020) (*Cum Laude*)

72. E. Dicks: *Grey matter covariance networks in Alzheimer's disease: Edging towards a better understanding of disease progression* (09-09-2020)

73. J.J. van der Zande: *A sharper image of dementia with Lewy bodies: the role of imaging and neurophysiology in DLB, and the influence of concomitant Alzheimer's disease pathology* (21-09-2020)

74. I. van Steenoven: *Cerebrospinal fluid biomarkers in dementia with Lewy bodies – towards a biological diagnosis* (22-09-2020)

75. N. Beker: *Cognition in Centenarians - evaluation of cognitive health and underlying factors in centenarians from the 100-plus Study* (02-10-2020)

76. F. de Leeuw: *Nutrition and metabolic profiles in Alzheimer's disease* (03-12-2020)

77. T. Timmers: *Tau PET across the Alzheimer's disease continuum* (02-12-2020)

78. E.E. Wolters: *Untangling tau pathology using PET* (02-12-2020)

79. A de Wilde: *Visualizing brain amyloid-beta pathology: Toward implementation of amyloid imaging in daily memory clinic practice* (17-3-2021)

80. C. Groot: *Heterogeneity in Alzheimer's Disease: A Multi-Modal Perspective* (6-4-2021) (*Cum Laude*)

81. D. Bertens: *The use of biomarkers in non-demented patients for clinical trial design and clinical practice* (14-4-2021)

82. L.E. Collij: *The AMYPAD project: Towards the next stage in amyloid PET imaging* (1-7-2021)

83. C.T. Briels: *Evaluation and implementation of functional cerebral biomarkers in Alzheimer's disease* (15-9-2021)

84. N. Tesi: *The Genetics of Cognitively Healthy Centenarians* (28-9-2021)

85. L.M. Reus: *Triangulating heterogeneity in dementia: An interaction between genetic variation, biological correlates and clinical symptoms* (28-9-2021)